

N. IZYUMOV, D. LINDE

**FUNDAMENTALS
OF
RADIO**

**MIR PUBLISHERS
MOSCOW**



Н. М. ИЗЮМОВ,
Д. П. ЛИНДЕ

Основы радиотехники

Издательство «Энергия»
Москва

N. IZYUMOV, D. LINDE

FUNDAMENTALS OF RADIO

Translated from the Russian
by
A. ULYANOV

**MIR PUBLISHERS
MOSCOW**

First published 1976

Revised from the 1971 third Russian edition

The Russian Alphabet and Transliteration

Аа	a	Кк	k	Хх	kh
Бб	b	Лл	l	Цц	ts
Вв	v	Мм	m	Чч	ch
Гг	g	Нн	n	Шш	sh
Дд	d	Оо	o	Щщ	shc
Ее	e	Пп	p	Ъ	"
Ёё	ë	Рр	r	Ы	y
Жж	zh	Сс	s	Ь	ʹ
Зз	z	Тт	t	Ээ	e
Ии	i	Уу	u	Юю	yu
Йй	y	Фф	f	Яя	ya

The Greek Alphabet

Αα	Alpha	Ιι	Iota	Ρρ	Rho
Ββ	Beta	Κκ	Kappa	Σσ	Sigma
Γγ	Gamma	Λλ	Lambda	Ττ	Tau
Δδ	Delta	Μμ	Mu	Υυ	Upsilon
Εε	Epsilon	Νν	Nu	Φφ	Phi
Ζζ	Zeta	Ξξ	Xi	Χχ	Chi
Ηη	Eta	Οο	Omicron	Ψψ	Psi
Θθ	Theta	Ππ	Pi	Ωω	Omega

На английском языке

CONTENTS

Foreword	9
Chapter I. PRINCIPLES OF RADIO COMMUNICATION . . .	11
1-1. Basic Properties of Electromagnetic Waves	11
1-2. General Principles of Radio Communication	15
1-3. Electromagnetic Waves Used in Radio Communication . . .	18
1-4. A Brief History of Radio in the USSR	20
1-5. Components Used in Radio Engineering	23
Chapter II. AC CIRCUITS	26
2-1. Sinusoidal Quantities and Their Vector Representation . .	26
2-2. Basic Components of Radio Circuits and Alternating Currents in Them	28
2-3. AC Power	43
2-4. Steady-State and Transient Processes in Electric Circuits Including Capacitors or Coils	48
2-5. Nonsinusoidal Currents and Their Spectra	55
2-6. Free Oscillations in a Circuit	63
2-7. Forced Oscillations in a Series-Connected Circuit	66
2-8. Forced Oscillations in a Parallel-Connected Circuit	72
2-9. Transient Processes in Oscillatory Circuits	81
2-10. Parallel Circuits with Reactive Elements of Both Types in One of Their Branches	86
Chapter III. COUPLED CIRCUITS	92
3-1. Oscillations in Two Coupled Circuits	92
3-2. Tuning a System of Two Coupled Circuits	104
Chapter IV. ELECTRICAL FILTERS	106
4-1. Purpose of Filters	106
4-2. Filters for DC Supplies	106
4-3. Low-Pass Filters	109
4-4. High-Pass Filters	114

4-5. Bandpass and Band-Elimination Filters	117
Chapter V. TRANSMISSION LINES	120
5-1. Electrical Signals in Ideal Infinitely Long Lines	120
5-2. Signals in Finite Length Lines with Far End Open-Circuited	128
5-3. Signals in Finite Length Lines with Far End Short-Circuited	134
5-4. Signals in Lines with a Reactive Load	136
5-5. Signals in Lines with a Resistive Load	138
5-6. Signals in Lines with a Combined Load	143
5-7. Actual Lines with Losses	145
5-8. Transmission Lines as Reactive Elements and Impedance Transformers	161
Chapter VI. ANTENNAS	169
6-1. Radiating Systems	169
6-2. Double-Dipole Antennas	172
6-3. Effect of the Ground on Antenna Radiation. Asymmetrical Dipoles	187
6-4. Antenna Resonant Frequencies. Harmonic Antennas	192
6-5. Inphased and Antiphased Antennas. Reflectors and Directors	195
6-6. Ground Effect on Antenna Radiation Patterns	203
6-7. Complex Dipoles	209
6-8. Loop Antennas	212
6-9. Long- and Medium-Wave Antennas	214
6-10. Short-Wave Antennas	217
6-11. Ultrashort-Wave Antennas	225
Chapter VII. RADIO WAVE PROPAGATION	233
7-1. Properties of Atmosphere and Ground Affecting Radio Wave Propagation	233
7-2. Radio Waves Propagating in Atmosphere. General Regularities	245
7-3. Long-Wave Propagation	252
7-4. Medium-Wave Propagation	253
7-5. Short-Wave Propagation	256
7-6. Ultrashort-Wave Propagation	265
7-7. Electromagnetic Waves in Outer Space	272
Chapter VIII. VACUUM AND SEMICONDUCTOR DEVICES	277
8-1. Modern Electronics	277
8-2. Motion of Electrons in Vacuum. Cathodes of Electron Valves	279
8-3. Diodes	290
8-4. Triodes	302
8-5. Multigrid Electron Valves	324
8-6. Electric Conduction in Semiconductors	337
8-7. P-N Junction and Crystal Diodes	343
8-8. Transistors	351

8-9. Miniaturization of Electronic Devices	363
8-10. Cathode-Ray Tubes	365
Chapter IX. PRIMARY-SIGNAL AMPLIFIERS	372
9-1. Purpose and Classification	372
9-2. Audio-Frequency Amplifiers. General	378
9-3. Audio-Frequency Small-Signal Amplifiers	393
9-4. Audio-Frequency Output Amplifiers	403
9-5. Driver Stages. Feedback in Amplifiers	415
9-6. Video Amplifiers	421
Chapter X. WAVE GENERATION	428
10-1. Operating Principles of Valve Oscillators	428
10-2. Separately Excited Oscillators (Amplifiers)	435
10-3. Self-Excited Oscillators	449
10-4. Ultrahigh-Frequency Valve Oscillators	473
10-5. Klystron Amplifiers and Oscillators	481
10-6. Travelling-Wave Oscillators	488
10-7. Backward-Wave Oscillators	492
10-8. M-type Travelling-Wave Oscillators	496
10-9. Transistor Oscillators and Amplifiers	507
10-10. Negative-Resistance Oscillators	510
10-11. Sinewave RC Oscillators	511
10-12. Frequency Pulling in Self-Oscillators	513
10-13. Self-Oscillator Lock-in	517
10-14. Nonsinewave Oscillators	523
Chapter XI. CONVERSION OF ELECTRIC SIGNALS	532
11-1. Concept of Signal Conversion	532
11-2. Amplitude Modulation	533
11-3. Frequency and Phase Modulation	545
11-4. Pulse Modulation	555
11-5. Detection of Radio Signals	561
11-6. Frequency Converters	565
11-7. Conversion of Electric Pulses	571
Chapter XII. RECEIVERS	579
12-1. Purpose and Basic Characteristics	579
12-2. Receiver Input Circuits	587
12-3. High-Frequency Amplifiers	595
12-4. Intermediate-Frequency Amplifiers	607
12-5. Radio Interference	611
12-6. Frequency Converters	618
12-7. Receiver Detector Stages	633
12-8. Controls and Adjustments in Receivers	647
12-9. Examples of Receiver Circuitry	653
Reference Data	659
Index	664

FOREWORD

These days radio engineering has become a very important branch of science solving a large number of problems associated with economic, technological and cultural progress. Every year, it finds ever increasing application and the number of people using radio equipment constantly grows. Many of these people have only rudimentary or no knowledge of radio engineering, although modern radio equipment is often so complicated that its effective use is impossible without some training.

The wide sphere of radio application in different branches of science and technology, as well as its close connexion with art and sport has also created a great number of radio amateurs in all countries. Some build radio receivers, tape recorders and TV sets, others design radio controlled models, short and ultrashortwave transmitters or equipment for a fascinating game called "hunting for a fox", etc.

All this increases the interest in radio engineering knowledge on the part of an ever growing number of people. The study of radio is, however, made more difficult for the majority of readers as it is usually explained with the use of higher mathematics. On the other hand, when higher mathematics is not used, many important problems are often oversimplified and treated without sufficient explanation and demonstration. Moreover radio engineering is a coherent science in which everything is interrelated and interdependent; therefore lack of understanding of fundamental phenomena and laws prevents the reader from fully understanding further problems.

It is far from clear what one should understand under the name of "radio engineering" and its fundamentals since this

branch has been extended, diversified and become interwoven with many other branches of science and technology. Under "radio engineering" proper one usually understands the use of electromagnetic radiation for the obtaining of information from a distant source. This is effected through the use of a transmitting (radiating) device and a receiving device provided conditions for propagation of radio waves are favourable. In accordance with this, the book describes the operating principles of radio transmitters, radio receivers and radiating devices, as well as radio wave propagation. It goes without saying that one book cannot exhaustively deal with all the varieties of existing radio circuits and devices; therefore, we concentrate our attention only on the most important and representative types.

Chapters 7, 8, 9 and 12 were written by N. M. Izyumov; the rest, by D. P. Linde.

Chapter I. PRINCIPLES OF RADIO COMMUNICATION

1-1. BASIC PROPERTIES OF ELECTROMAGNETIC WAVES

By the end of the 19th century scientists have discovered and investigated the properties of invisible electromagnetic waves capable of propagating over vast distances without appreciable attenuation. These waves were called *radio waves*, the term "radio" originates from the word "radiation".

On the basis of empirical data obtained by various naturalists, the British physicist Maxwell developed in the 1860s the electromagnetic theory and established the general nature of radio and light waves, as well as the laws of their propagation. Later, other kinds of radiation, such as ultra violet, infrared, X-rays, etc., were investigated. In spite of the differences between these kinds of radiation, they were shown to be identical in that they are all electromagnetic waves, the differences in their physical characteristics being determined by their wavelengths.

In 1886-1888 H. Hertz proved experimentally the main conclusions of Maxwell's theory by demonstrating that the laws of propagation, reflection and refraction of radio waves are similar to those of light.

When formulating the electromagnetic theory of light Maxwell was at once confronted with the difficulty of explaining its propagation as all the wavelike motions of matter known hitherto were associated with mechanical motions or elastic interaction of particles in that particular media where such motions took place. Waves on the surface of water, for example, travel due to the forces of internal friction and surface tension of water while sound propagates due to the elastic

deformation of a particular media or oscillations of gas molecules, etc. Thus, in vacuum these oscillations cannot propagate.

How should the unhindered propagation of light waves in space which is considered to be almost ideal vacuum be then explained? Maxwell suggested that all space was filled with some imperceptible kind of matter which he called ether, and that the propagation of electromagnetic waves, including light, was caused by oscillations of ether particles.

This motion or displacement of ether particles was given the name of displacement current. Indeed, if one places two plates inside of a vacuum vessel and connects them to a source of alternating emf, a magnetic needle placed nearby will be influenced by the alternating magnetic field in exactly the same way as if electrons were flowing in the gap between these electrodes. This current is commonly called *convection current*.

Although experiments fully and quantitatively corroborated the theory, the phenomena could not be qualitatively explained by this mechanical model.

Further attempts to discover ether and reveal its properties by means of the most sophisticated experiments not only failed to yield any results but proved that if ether were supposed to exist it would be impossible to explain many physical phenomena. Nowadays this concept is devoid of any physical sense, although this term is still habitually used.

Meanwhile physics continued to obtain new experimental data about the properties of electromagnetic waves. The remarkable experiments made by P. N. Lebedev in 1901 allowed him to discover and measure the pressure of light. Later it was proved that the particle emitting electromagnetic waves loses part of its mass. Finally, the study of elementary particles and their behaviour showed that they can under certain conditions be transformed into electromagnetic radiation while electromagnetic radiation can sometimes turn into charged particles. It was discovered that electrons, on the one hand, behave as elementary particles of matter and, on the other, they exhibit certain wave properties such as ability to diffract, that is, pass around obstacles. Electromagnetic radiation, in its turn, possesses corpuscular properties associated with discontinuities. In other words, it behaves as a flow of minute particles.

All these facts led to the conclusion that electromagnetic waves are a special form of moving matter.

Apart from the ether hypothesis, Maxwell's electromagnetic theory gives a correct representation of particular physical reality, generalizing empirically established fundamental laws of electricity. An important conclusion of the theory is that an alternating electric field produces an alternating magnetic field. From the law of electromagnetic induction it follows that alternating electric and magnetic fields always co-exist and are quantitatively interrelated. An alternating electric field always produces an alternating magnetic field and vice versa. Accordingly, every disturbance, that is, change in electric or magnetic field results in a combined alternating electromagnetic field.

An alternating-current electromagnetic field has a most important property. From the point where it occurs the disturbance begins to propagate in all directions producing electromagnetic waves, rather than to be localized in this point. A similar effect one can observe on the surface of a lake when a stone thrown into water produces spreading waves.

Electromagnetic waves are carriers of energy. Life on the Earth owes its existence to the radiant energy brought by them from the surface of the Sun. It should be clear then that electromagnetic energy generation must be the result of conversion of energy. Such conversion occurs, for instance, when an electron is accelerated. The energy spent by the device accelerating the electron is converted into the energy of electromagnetic radiation.

It is known from physics that electrons orbit around the positively charged atomic nucleus, each having a constant energy relative to the nucleus. Under certain conditions electrons can jump from one orbit into another with their energy being changed and the resulting excess of energy transferring to electromagnetic radiation. This method of electromagnetic wave generation is used in quantum electronics, at present a rapidly developing branch of science and technology.

In radio engineering accelerated free electrons found in vast amounts in conductors are commonly used.

It is evident that in practice electrons cannot be accelerated in one direction for a long time. Therefore it becomes necessary to alternate acceleration with deceleration, i.e. acceleration

in the opposite direction. This cycle can be repeated many times by connecting some conductors to the terminals of a source of an alternating emf which causes electrons in the conductors to oscillate, thus producing electromagnetic radiation in the surrounding space. Such a conductor converting the energy of an alternating emf into that of electromagnetic waves is called an *antenna* (aerial).

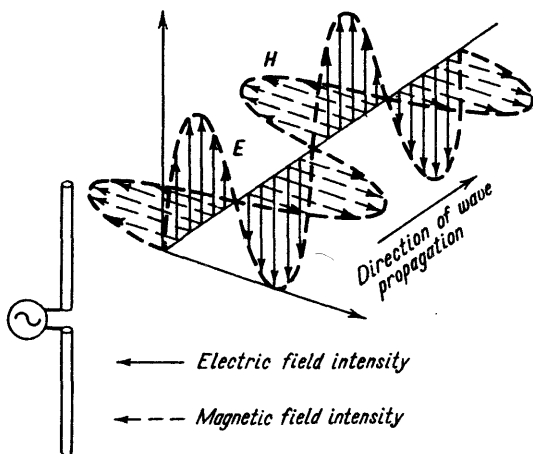


Fig. 1-1. Structure of electromagnetic-wave field

On the basis of Maxwell's theory it was found that the velocity of propagation of electromagnetic waves in any medium is

$$v = \frac{c}{\sqrt{\epsilon\mu}} \quad (1-1)$$

where c is the velocity of light in free space; ϵ is the dielectric constant of the medium, and μ is the magnetic permeability of the medium. For the air $\epsilon \approx \mu \approx 1$ and the propagation velocity of electromagnetic waves is close to that of light in free space

$$v = c \approx 300,000 \text{ km/s}$$

The oscillations of electrons in the antenna are produced by a source of a periodically changing emf having a period T .

If, at some moment of time, the instantaneous intensity of the electromagnetic field at the antenna is a maximum (Fig. 1-1), it will reach the same value after the time period T . The distance travelled during this time by the electromagnetic field produced by the antenna at the initial moment will be

$$\lambda = vT \quad (1-2)$$

The minimum distance between two successive points in space where the field has the identical magnitude is called *wavelength*. As follows from expression (1-2), the wavelength λ depends on the propagation velocity and the oscillation period of electrons in the antenna. Since the resulting current changes with frequency $f=1/T$, the wavelength is

$$\lambda = \frac{v}{f} \quad (1-3)$$

On the basis of Maxwell's theory, as well as from experiments it can be seen that electromagnetic waves are of transversal type. This means that both the magnetic and electric field vectors of the electromagnetic wave are everywhere perpendicular to the direction of its propagation at all times, the direction of the electric field being always perpendicular to that of the magnetic field (see Fig. 1-1).

1-2. GENERAL PRINCIPLES OF RADIO COMMUNICATION

The following question may arise. Is it possible to transmit human speech by means of radio waves in such a way as to convert sound waves into electric oscillations which, in turn, will be converted into electromagnetic waves in an antenna, so that these waves at the point of reception will again be converted into sound waves?

The oscillations produced by human voice are of low frequency ranging from 75 to 3000 cycles per second (Hz). Using formula (1-3) it can be easily shown that such oscillations correspond to radio waves from 4000 to 100 km in length. Antenna, however, can effectively radiate electromagnetic waves only if its size is close to the length of the wave to be transmitted. This means that the transmission of oscillations having such wavelengths is practically impossible.

From the fact that antennas can be built with dimensions not exceeding a few hundreds of meters, it becomes clear that only wavelengths of up to several kilometers are suitable for radio communication. Such electromagnetic waves are produced by oscillations with frequencies much higher than those of sound signals. The former, therefore, can only be used as carriers of useful signals. This is done by varying the amplitude,

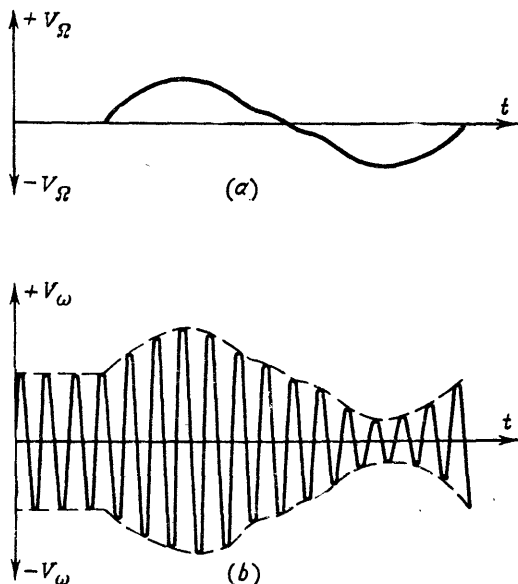


Fig. 1-2. Signals used in radio transmission

(a) low-frequency signal; (b) high-frequency signal modulated in amplitude by low-frequency signal

frequency or phase of carrier waves (carriers) in accordance with the sound signal. It is also possible to code the message to be transmitted in the form of certain combinations of pulses varying in frequency, duration and mutual position. The process of changing the above parameters of carriers is called *modulation*.

Figure 1-2 shows a sound signal and amplitude-modulated wave which results when the high-frequency oscillations are

changed in amplitude in accordance with the signal being transmitted.

Thus, any radio-transmitting system comprises three main elements: a source of alternating emf, i.e. oscillator, a device where modulation takes place, i.e. modulator, and an antenna.

The reception point must include some device, a receiving antenna, which transforms the energy of electromagnetic waves into that of electrical oscillations. Electromagnetic waves sent by many transmitters operating at various frequencies come to this antenna. In order to receive signals from

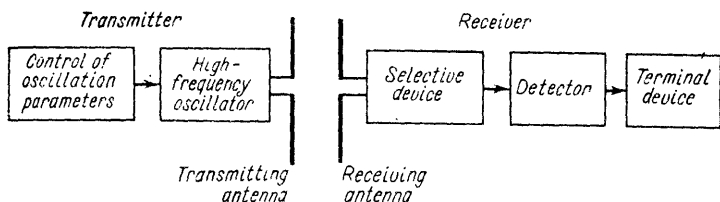


Fig. 1-3. Block diagram of radio communication system

one particular radio station, it is necessary to have a *selective device* which can separate the desired signal frequency from all other signal frequencies. To solve this problem, electrical oscillatory circuits tuned to the frequency of the desired radio station are used. The high-frequency oscillations obtained from such selective circuit must be then converted into the currents or voltages changing in accordance with the modulating signals in the transmitter. To accomplish this, the receiver must include a special device known as a *detector*.

Finally, the detected signal must be fed to some *terminal device* which will record it or enable a human being to perceive it in the form of sound or light. Thus, the receiver must comprise the following elements: an antenna, a selective device, a detector, and a terminal device.

A general block diagram of a radio communication system is shown in Fig. 1-3.

1-3. ELECTROMAGNETIC WAVES USED IN RADIO COMMUNICATION

Research has shown that the conditions of propagation of radio waves considerably depend on their length. This is due to differences in the propagation of two waves radiated by the antenna: one along the surface of the Earth (called the *ground* or *surface wave*) and the other, at some angle to the horizon (called the *sky wave*).

These two waves are shown in Fig. 1-4.

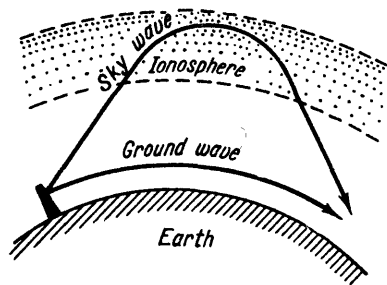


Fig. 1-4. Paths of radio-wave propagation

The ground wave travelling over the Earth's surface induces alternating currents in it, which will suffer the greater attenuation the higher is the operating frequency. Therefore, absorption of the energy of the ground wave tends to increase with

an increase in operating frequency.

The sky wave, however, reaches the highly ionized upper layers of the atmosphere (the ionosphere). The incoming electromagnetic wave sets ions and free electrons in these ionized layers in an ordered oscillatory motion which in turn results in radiation. Some of this radiation can return to the ground. Not all the incoming wave energy is transferred to the reflected wave since it is partially converted into heat. Both theory and practice have shown that the amount of energy being absorbed increases with the increase in wavelength (i.e. with the decrease in frequency).

For waves with wavelengths above 3000 meters attenuation in ground is very low, whereas attenuation in ionosphere is very high. Moreover, owing to the fact that they are effectively guided around the Earth's surface, the ground waves can provide radio communication over distances up to several thousands of kilometers. The sky wave, on the other hand, cannot be used for communication because of high attenuation in the ionosphere.

Waves with lengths exceeding 1000 meters are called *long waves* (corresponding to the frequencies less than 300 kilohertz).

For shorter waves (from 1000 to 100 meters) called *medium waves* the attenuation of the ground wave increases and radio communication at the corresponding frequencies (from 300 kilohertz to 3 megahertz) can be effected only over comparatively short distances. However, the sky wave reflected by the ionosphere has at these frequencies the intensity sufficient to maintain communication over distances up to several thousands of kilometers.

For the waves with lengths from 100 to 10 meters known as *short waves* (frequencies from 3 to 30 megahertz), the attenuation in the ground becomes so high that the ground wave fails to provide communication over distances greater than some dozens of kilometers. On the other hand, the attenuation of the sky wave in the ionosphere at these frequencies sharply decreases, and it can maintain communication over great distances.

Still shorter waves (less than 10 meters), called *ultrashort waves* (frequencies above 30 megahertz), usually pass through the ionosphere without being reflected. As radio waves become shorter, their properties approximate more and more those of light waves.

Ultrashort waves almost lack the ability to follow the Earth's surface, hence they can provide communication only over distances limited by the line of sight. This, of course, limits the use of ultrashort waves, yet on the other hand, precludes mutual interference between radio stations located within an area of more than several dozens of kilometers from one another. Moreover, this frequency range is practically free from atmospheric and industrial interference. It should be noted, however, that the lower atmospheric layers under special conditions can refract ultrashort waves so that they can also return to the ground. Continuing studies of this phenomenon offer new possibilities for a still wider use of ultrashort waves for telecommunication and long-distance TV reception.

The use of ultrashort waves has opened up a number of new interesting possibilities to radio engineering. For example, owing to their small wavelength it is possible to build

sophisticated antenna systems which can form narrow beams in only required directions. Radars, TV broadcast stations, remote control systems and many other special systems can be effected only in this frequency range. Therefore this frequency range has recently become the focus of attention for radio engineers. At present even the waves as short as several millimeters find application in practice.

It is necessary to point out that in literature one can often encounter a different radio wave range division. In accordance with the earlier International agreement waves longer than 3 km were referred to as long, those between 3 km and 200 m, as medium, the ones between 200 and 50 m, as intermediate, and those from 50 to 10 m, as short waves.

1-4. A BRIEF HISTORY OF RADIO IN THE USSR

A wireless message—the first one in history—was sent over a distance of about 250 meters on March 24, 1896 with the radio communication system built by A. S. Popov.

In Popov's transmitter (Fig. 1-5), a source of direct current was connected to a convertor which produced a high output voltage. This voltage was applied to two relatively long wires separated by a spark gap; the wires served as an aerial. When the voltage across the aerial wires exceeded the breakdown voltage of the spark gap, a spark discharge took place and damped electrical oscillations were induced in the aerial circuit. These oscillations set up in the surrounding space an electromagnetic radiation which was controlled with the help of a telegraph key switching on and off the convertor supply. This made it possible to use the Morse code in producing electromagnetic signals.

Popov's receiver consisted of a receiving aerial whose circuit included a coherer, a glass tube filled with metal powdered filings. When alternating current appeared in the circuit of the receiving aerial, the individual particles of the metal powder coalesced together, resulting in a sharp fall in the coherer resistance. Thus, the coherer closed the circuit of an electric bell or a relay winding of a recorder in a telegraph set. When the coherer resistance decreased, a current flowed in the bell circuit, the bell hammer struck the sounding cup and the telegraph signal could be heard or recorded. During

the return movement the hammer struck the coherer, thus restoring its initial (high resistance) state. Here the coherer had the function of a detector, and the bell, that of a terminal device.

A receiver of this type was first demonstrated by A. S. Popov on May 7, 1895 at the session of the Russian Physical and

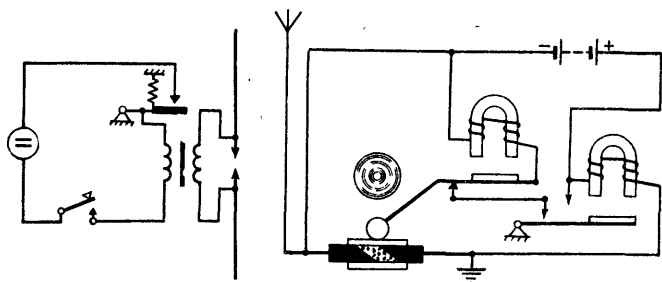


Fig. 1-5. A. S. Popov's transmitter and receiver

Chemical Society in St. Petersburg. This day is now considered to be the birthday of radio.

A. S. Popov was constantly improving his system; the first message, as mentioned above, was sent over a distance as short as 250 meters, but only a year and a half later radiograms could be transmitted over several kilometers, and in 1899, during the salvaging of the "Admiral Apraksin" battleship, Popov maintained radio communication over a distance of 40 km. In 1901, using Popov's station, radio contact was established over a distance of 150 km.

Throughout the world Popov's invention was widely recognized. Spark telegraph stations sprang up everywhere, and only in Russia, because of the inertia and complete indifference of the czarist government, was the development of radio greatly hindered.

In 1906, the American inventor Lee de Forest built the first three-electrode vacuum tube which opened up new vistas for radio engineering. In 1913, the German engineer Meissner built the world's first valve oscillator. The invented valve oscillators and sustained oscillation amplifiers solved the problem of building compact, stable, tunable transmitters

and receivers which enabled not only telegraph signals, but also human voice, music and pictures to be broadcasted.

From the first days of Soviet power, the Communist party led by V. I. Lenin gave great attention to the development of radio. Under instructions by Lenin, a radio laboratory was set up in Nizhni Novgorod where the most important work concerning the designing of radio equipment was concentrated. Under the hard working conditions of the time, starvation, blockade and economic dislocation, the laboratory staff headed by M. A. Bonch-Bruevich designed and built prototypes of home-made electronic valves, and as early as 1920 a rather powerful radio telephone station was constructed near Moscow. Two years later the world's most powerful broadcasting station was built in Moscow. Since that time the Soviet Union holds the first place in the world as to the power of its radio broadcasting stations.

In the early 1920s high-power and superpower oscillator valves of up to 100 kW were developed in the USSR under the guidance of M. A. Bonch-Bruevich. In 1933 a radio station with the radiated power of 500 kW developed by a team of engineers headed by A. L. Mintz, was built near Moscow. Further high-power stations were constructed in later years.

Nowadays all big cities in the USSR have trunk-line service. Local radio stations broadcast their own programmes in addition to relaying central radio broadcast programmes. In addition to high-power radio broadcast stations, dozens of thousands of various special-purpose radio stations are in operation in aeroplanes, trains, ships, etc. There are also thousands of radio amateurs who use short-wave radio sets for long-distance contacts, "fox-hunters" (a special kind of sport), amateur designers constructing radio-controlled models of aeroplanes and ships as well as many other types of different purpose radio equipment.

In the late 1930s and especially after World War II other branches of radio engineering developed rapidly: television, remote control of different equipment by means of radio (telecontrol), radiolocation (radar), radio navigation, etc.

The methods of radio engineering are now used in various fields of science and technology, e.g. in physics, chemistry, geology, medicine, astronomy, mathematics, etc. At present there is no branch of science where use is not made of some

kind of radio equipment. Distant areas of the Universe are studied with the help of radio. Spacecraft are guided by radio. Radio devices have made it possible the information to be obtained about the mysterious and amazing phenomena taking place in far-away Galaxies as well as inside atomic nuclei. Therefore some acquaintance with the fundamentals of radio, which is both useful and fascinating, is now indispensable for wide circles of people in the most diverse professions.

1-5. COMPONENTS USED IN RADIO ENGINEERING

In spite of the great variety of functions performed by different radio devices they all consist generally of a limited class of components such as valves, transistors, resistors, coils and capacitors. Therefore, when studying radio engineering one should begin with the principles of operation of these components.

Assemblies obtained by interconnecting the above elements possess basically new properties. One of such basic assemblies is the oscillatory circuit consisting of a coil (inductor) and a capacitor, and forming part of a vast majority of generators and amplifiers used in transmitters and receivers.

Such a circuit is never used as an isolated unit; it is always connected or coupled with other circuits of a given transmitter or receiver. Interconnection of these circuits can alter considerably the properties of the resulting oscillatory system. Such interconnected circuits are called coupled circuits.

As can be seen from the block diagrams of the transmitter and receiver, currents of widely different frequencies flow in their circuits. Consequently, it often becomes necessary to filter out some frequencies, i.e. to pass the current of one frequency into a particular circuit while rejecting the others. This is done with the help of special circuits known as electric filters which consist of resistors, capacitors, coils and, sometimes, oscillatory circuits.

Radio equipment often includes systems of electric conductors which transmit high-frequency energy from one part of the equipment to another, for instance, from the transmitter to the antenna or from the antenna to the receiver. Such systems are called feeders. They must transmit the energy possibly without losses. When the length of the feeders becomes

commensurable with the operating wavelength, they exhibit a number of interesting characteristics, making it possible to employ them as inductances, capacitances, oscillatory circuits, filters, etc.

The study of feeder properties facilitates the understanding of many phenomena which occur in the antennas, which mostly represent systems of long wires. Depending on their shape, height above the ground and the methods of coupling with the feeders, the antennas possess different radiating properties. The necessity of constructing antennas with different characteristics is determined by the purpose of a given radio station and the conditions of wave propagation which depend on the frequency range used: some radio waves can propagate only within the line of sight, others can cover thousands of kilometers, and there are radio waves that can travel around the Earth several times following its surface. To select correctly the operating frequency and the antenna system, it is necessary to know the corresponding conditions of radio wave propagation.

At present, most amplifiers, generators and other basic units use electron valves or transistors as their main elements. These elements have the following important feature. A change in the voltage across one of the element electrodes changes the current passing in the circuit of another electrode to a much greater extent than a change in the voltage across the latter does. This makes it possible to obtain considerable changes of voltage across the second electrode by slightly changing the voltage across the first one, i.e. the amplification of alternating voltages.

These same devices can be used to provide modulation, detection and other forms of conversion. At present many types of valves and semiconductors performing different functions are available.

The knowledge of the operating principles of valves and oscillatory circuits allows one to study those of amplifiers and generators of which these two elements are usually basic parts. The invention of valve generators and amplifiers made it possible to build receivers and transmitters which can amplify signals millions of times.

From the general description of the communication system structure given above it can be concluded that in every radio

system, in addition to generation and amplification, various forms of conversion of oscillations such as frequency, amplitude or phase conversions take place. In modern radio equipment, this is usually done in valve or semiconductor circuits.

It should be emphasized that radio receivers and transmitters cannot be regarded simply as a mechanical assembly of the above-mentioned components. As was the case with the oscillatory circuit consisting of a coil and a capacitor, or with the amplifier including a circuit and a valve, the assembled system acquires new properties. In order to understand the functions of the whole system and those of its individual components one must know these new properties.

Knowledge of the principles underlying the construction and operation of radio transmitters and receivers enables understanding of the essence of new and highly important branches of radio engineering such as television, radiolocation, remote control, etc. Radio systems used for the detection of planes and ships for remote control of different equipment, electronic computers and hundreds of other special devices employ various pulse circuits the main of which are described below.

From this short introduction it is clear that radio engineering is a systematized and logical science in which all concepts are interconnected and inferred one from another. If you fail to understand something that precedes, you will not understand properly what follows. Therefore, readers are recommended to adhere to the exposition sequence as arranged in this book.

Chapter 11. AC CIRCUITS

2-1. SINUSOIDAL QUANTITIES AND THEIR VECTOR REPRESENTATION

Both direct and alternating currents flow in radio circuits. A current whose value and direction change with time is called *alternating current*. Variations of the alternating current can follow different laws. However, it usually changes with time according to the sine law (in this case it is also called *harmonic current*)

$$i = I_{\max} \sin \omega t \quad (2-1)$$

where i is the instantaneous value of the current; $I_{\max}$ denotes the amplitude, i.e. the maximum value of the current; ω is the angular frequency.

Every quantity which varies sinusoidally with time can be represented as a projection of a certain rotating vector on the vertical axis (Fig. 2-1). In one complete revolution or cycle of the vector, the sinusoidal quantity will twice reach its maximum. This occurs when the vector coincides with the vertical axis so that its projection is equal to the vector magnitude. The frequency of vector rotation ω is the angular frequency of the sinusoidal quantity. It is measured by the angle through which the vector turns in one second (usually in rad per 1 s). The angle by which the vector turns relative to the horizontal axis is called the *phase* of the sinusoidal quantity.

The time of one complete revolution of the vector corresponds to the period T of the sinusoidal quantity. The inverse value of the period gives the number of revolutions the vector makes per second. This is called the *frequency* of the sinusoidal quantity and is measured in hertz (Hz).

$$f = \frac{1}{T} = \frac{\omega}{2\pi} \quad (2-2)$$

For the frequency of 1 Hz, the vector completes one revolution per one second. If at the initial moment rotating vector I' is at an angle φ_0 relative to the horizontal axis, the sinusoidal quantity has the *initial* phase φ_0 . The sine wave with this initial phase proves to be shifted (in time) to the left relative to the sine wave with the zero initial phase (the dashed line in Fig. 2-1), i.e. *leading* the latter by the angle φ_0 .

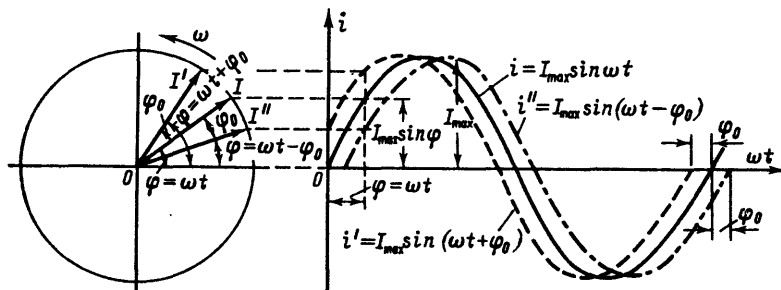


Fig. 2-1. Vector representation of a sinusoidal quantity

The equation of such a sine wave will then have the following form

$$i' = I_{\max} \sin(\omega t + \varphi_0) \quad (2-3)$$

The projection of the vector I'' placed on the other side of the vector I , which rotates at the same speed, represents a sine wave *lagging* behind the first one by the angle φ_0 (the dot-and-dash line in Fig. 2-1). Its equation is as follows:

$$i'' = I_{\max} \sin(\omega t - \varphi_0) \quad (2-3a)$$

In radio engineering many variables turn out to be proportional to the rate of change of one or another sinusoidal quantity. The speed of the rotating vector I is also a vector K tangent to the circle circumscribed by the vector I (Fig. 2-2). For the sake of convenience, let us represent the vector K as starting from the origin of the coordinate system (this will not change the vector projection on the vertical axis). It is easy to see that this vector leads the vector I by 90° .

From mechanics it is known that the magnitude of the linear velocity vector K is related to the angular velocity and

the magnitude of the rotating vector I by a simple relation

$$K = \omega I \quad (2-4)$$

By projecting this vector onto the vertical axis, we obtain the rate of change of the sinusoidal quantity, which also varies sinusoidally, leading the quantity itself by 90° .

From the curve which represents the sum of two sine waves which have the same frequency but differ in amplitude and

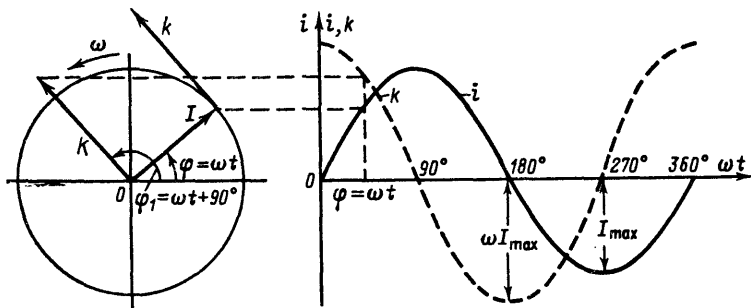


Fig. 2-2. Sinusoidal quantity and its rate of change

phase (Fig. 2-3), it can be seen that the resultant curve is a sine wave of the same frequency. It can be obtained as a projection of the vector C which is a sum of the vectors A and B whose projections are the sine waves to be added. The vector C is a diagonal of a parallelogram whose sides are the vectors A and B .

On the basis of the above property of rotating vector projections it is possible to find the result of adding sinusoidal quantities by geometrical summation of the corresponding vectors. A drawing showing the instantaneous positions of vectors gives amplitude and phase relations for variables in a circuit. Such a drawing is called *vector diagram*.

2-2. BASIC COMPONENTS OF RADIO CIRCUITS AND ALTERNATING CURRENTS IN THEM

Radio circuits comprise, basically, inductance coils, resistors and capacitors. Let us examine the operation of these components in AC circuits.

According to the electromagnetic induction law, any alternating magnetic field produces in a conductor placed in it an emf which is proportional to the rate of change of the magne-

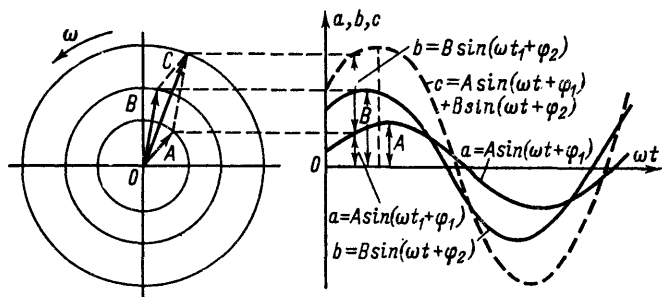


Fig. 2-3. Addition of two sinusoidal quantities

tic field. A conductor carrying a current is always surrounded by the lines of force of the magnetic field produced by the current (Fig. 2-4a). A change of the current in the conductor

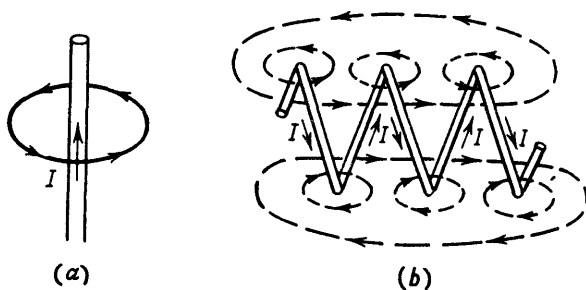


Fig. 2-4. Magnetic field produced by current flowing in a wire
(a) straight wire; (b) coiled wire

results in a proportional change of the magnetic field, which induces an emf in the conductor. This effect is known as *self-induction*. By Lenz's law, the induced emf tends to prevent the change that causes it. This emf is called emf of self-induction. Self-induction in electrical systems, which determines the circuit resistance to variations of its current, acts like inertia in mechanical systems.

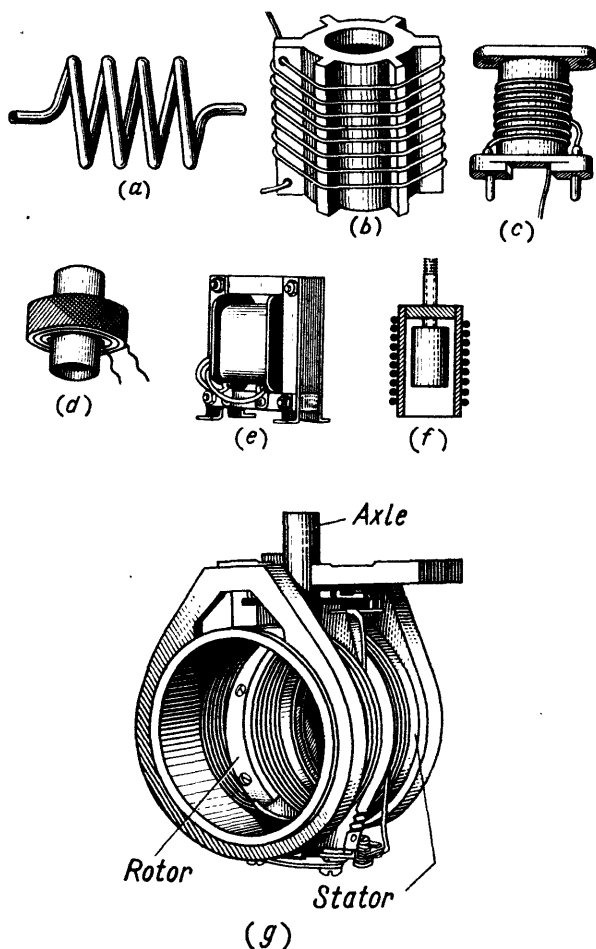


Fig. 2-5. Induction coils

(a) single-layer without former; (b) with ceramic former; (c) with plastic former; (d) multiple-layer; (e) choke with iron core; (f) with ferrite core; (g) variable-inductance coil (variometer)

We shall see later that it is often desirable to have a device which possesses the property of self-induction to a maximum degree. It is easy to understand that a coil wound of wire can serve as such a device (Fig. 2-4*b*). Every straight section of the wire is surrounded by the field induced only by the current flowing in it whereas the whole coil is threaded by the total field produced by all its turns. Therefore, the emf of self-induction produced in the coil appears to be much greater than that in the straight piece of wire of which the coil is wound.

Figure 2-5 shows representative types of coils used in radio engineering. For the case of low inductances, coils are wound in one layer on a special coil former or made of rigid wire without any former (Fig. 2-5*a* through *c*). To provide high inductances, multilayer coils of hundreds of turns are often used (Fig. 2-5*d*). If very high inductances are required, coils are wound on cores made of a ferromagnetic material (Fig. 2-5*e*). To make it possible to change the coil inductance, it is provided with a movable core (Fig. 2-5*f*). In order to vary inductance over a very wide range, a variometer is used (Fig. 2-5*g*). It consists of two coils connected in series, the stator and rotor; when the latter rotates, the total inductance is changed due to a change in the mutual inductance of the two coils.

The emf self-induced in a conductor carrying a current, produces the effect of an apparent rise in its resistance. This effect exhibits itself to a variable degree across the conductor cross section. This is easy to understand if one imagines the conductor as consisting of several tubular layers (see Fig. 2-6). The inner layer, which coincides with the axis, is surrounded by the magnetic field produced by the currents in all conductor layers. The tubular layer, next to the axis, is surrounded only by a part of the magnetic field since the remaining part of the field will be inside the tube. Thus the farther the layer from the axis, the lesser will be the self-induced emf in it, and the lesser will be its effective resistance to alternating current.

At very high frequencies (that is, at a very high rate of current change) the resistance of inner layers increases to such a degree that the bulk of the current flows through a very thin surface layer of the conductor. This is known as the *skin effect*. To obtain a low wire resistance at high frequencies, it is necessary to increase the diameter of the wire. For this pur-

pose tubular wires are sometimes used. Special attention should be given to the quality of the wire surface; therefore, the latter is often covered with a thin layer of silver, gold or copper which increases its conductivity.

On the contrary in some cases, elements with a high resistance are required; these are called *resistors* (Fig. 2-7). They

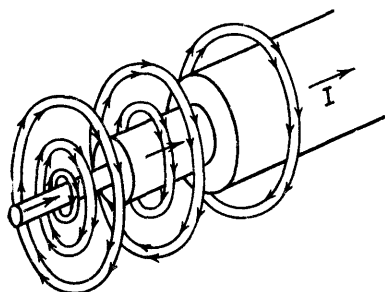


Fig. 2-6. Magnetic field surrounding different layers of conductor

can be made from wire having a high specific resistance, which is wound in a planar spiral or around a ceramic former. A very thin layer of special varnish or carbon with a low conductivity, deposited on the former, presents another type of resistors. Also, bulk resistors made from carbonized ceramic rods, exist.

In order to store electric energy a device known as a *capacitor* is used (Fig. 2-8). In the capacitor, electricity is stored due to the fact that two closely spaced conducting plates separated by a dielectric material retain opposite charges, thus allowing more charges to move to the plates.

It is clear that the less the spacing between the plates, the stronger will be the mutual attraction of the opposite charges; and the larger the plate area, the greater will be the number of charges. Therefore, capacitors are made of rolled sheets of thin metal foil separated by oiled paper, or as a set of metal plates separated by air gaps (Fig. 2-8a), mica sheets or other dielectric materials. A very thin dielectric layer is used in electrolytic capacitors (Fig. 2-8c) where the aluminium oxide is formed electrolytically on aluminium foil. Sometimes a very thin metal layer is deposited directly on lacquered paper which is then rolled and enclosed in a ceramic or metal tube (Fig. 2-8e). Capacitors with low capacitance values are often fabricated in the form of ceramic or metal tubes covered both from inside and outside with a layer of a conducting material. Variable capacitors (Fig. 2-8b) consist of a set of fixed plates (stator) and a set of rotating plates (rotor); when the latter rotates, the plates effective surface is changed. If capacitance

adjustments are seldom required, the rotor axle is provided with a screwdriver slot rather than with a knob hand (Fig. 2-8d). Capacitors of the latter type are referred to as trimmer capacitors.

Let us examine an alternating current flowing through a capacitor. The voltage (in volts) u_C across the capacitor terminals

$$u_C = \frac{q}{C}$$

where q is the charge on the capacitor (in coulombs) and C is its capacitance (in farads). If the voltage across the capaci-

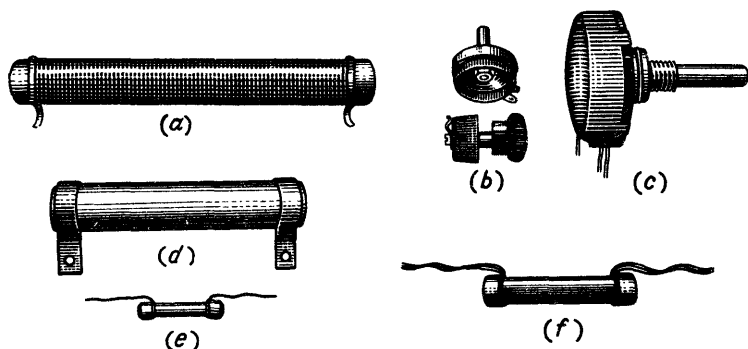


Fig. 2-7. Resistors

(a) vitrified wire-wound resistor; (b) wire-wound variable resistor; (c) deposited-carbon variable resistor; (d) ceramic resistor with carbon coating; (e) miniaturized resistor; (f) metal-film resistor

tor changes by Δu_C over the time interval Δt * this will cause the capacitor charge to change by the value

$$\Delta q = C \Delta u_C$$

Dividing both sides of the above equation by Δt we obtain

$$\frac{\Delta q}{\Delta t} = C \frac{\Delta u_C}{\Delta t}$$

* Here and elsewhere the Greek letter Δ (delta) denotes small differences or increments of the quantity which follows this symbol. For instance, $\Delta t = t_2 - t_1$, where t_1 and t_2 are two successive moments of time.

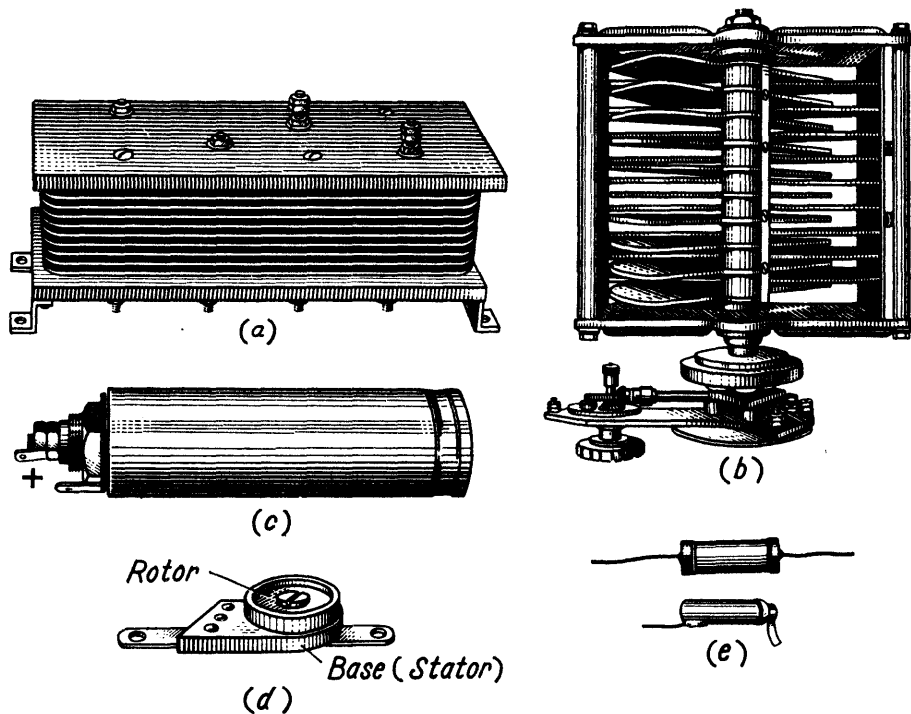


Fig. 2-8. Capacitors

(a) fixed air capacitor; (b) variable air capacitor; (c) electrolytic capacitor; (d) ceramic trimmer capacitor; (e) hermetically sealed paper capacitor

The left-hand side represents the mean value of the current in the capacitor circuit for the time interval Δt , and the ratio $\Delta u_C / \Delta t$ is the average rate of change of the voltage across the capacitor. Consequently, when the voltage across the capacitor changes, the current flowing in the capacitor circuit is equal to the product of the capacitance and the rate of change of the voltage.

If the voltage changes by the sine law, i.e.,

$$u_C = V_{\max C} \sin \omega t$$

then, on the basis of the general properties of the rate-of-change vector of a sinusoidal quantity discussed above, the current will also vary by the sine law and will lead the voltage by 90° .

$$i_C = C \frac{\Delta u_C}{\Delta t} = \omega C V_{\max C} \sin (\omega t + 90^\circ) \quad (2-5)$$

Let us explain the cause of this phase difference between the current and voltage. Suppose that at an initial moment the capacitor voltage u_C is maximum, this means that the capacitor is fully charged and the charging current i_C is zero. The capacitor voltage then slowly decreases (i.e. the capacitor is slowly discharged) and the discharge current is low. Gradually, the rate of voltage fall is increased and the discharge current increases. It reaches its maximum when the voltage becomes zero. Further the capacitor is charged with the charging current leading the voltage by 90° for the same reason (Fig. 2-9).

From expression (2-5) it follows that the current magnitude is

$$I_{\max C} = V_{\max C} \omega C \quad (2-6)$$

The ratio of voltage to current gives the effective opposition of the capacitor to alternating current. How should we then understand the opposition offered by the capacitor to alternating current in view of the fact that the insulator between its plates does not at all permit electrons to pass? When an alternating voltage is applied to the capacitor, its plates are successively charged and discharged, and an alternating current flows through the capacitor circuit as if the capacitor let it through, offering a certain finite resistance. During charging,

energy is stored by the capacitor in the form of the energy of an electrical field between the capacitor plates, and during discharge the capacitor gives the stored energy back into the circuit. Thus an ideal capacitor does not absorb energy for the full cycle, and the capacitor opposition to alternating current, unlike resistance of the wire, does not result in heat

losses. Therefore, the opposition offered by the capacitor to alternating current is called *capacitive reactance*.

The capacitive reactance is defined as

$$X_C = \frac{V_{\max} C}{I_{\max} C} = \frac{1}{\omega C} \quad (2-7)$$

This relation is explained as follows. For some fixed voltage a capacitor with a higher capacitance gives a greater current in its circuit. This is due to the fact that in order to increase the capacitor voltage a greater charge must be stored at its plates. For given capacitor voltage and capacitance, the higher the frequency, the greater will be the current (since at a higher frequency the capacitor must charge for a shorter interval of time); on the other hand, any increase in the capacitor current indicates a fall in its reactance (see Fig. 2-9).

In a certain sense, an induction coil in an AC circuit behaves in the opposite way. When an alternating current i_L flows through the coil

$$i_L = I_{\max} L \sin \omega t$$

an emf is self-induced in the coil, which is proportional to the rate of change of the current and is defined by the following equation

$$e_L = -L \frac{\Delta i_L}{\Delta t} \quad (2-8)$$

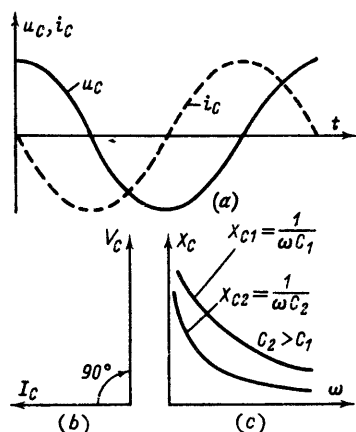


Fig. 2-9. Passing of alternating current through capacitor

(a) voltage across the capacitor and current in its circuit; (b) vector diagram; (c) capacitor reactances as functions of frequency

where L is the coil inductance (in henries). The minus sign in this formula accounts for the fact that the emf tends to prevent the change of current, i.e., when the current increases ($\Delta i_L > 0$) the emf is directed against it, and when the current decreases ($\Delta i_L < 0$) the emf acts in the same direction tending to maintain its initial value.

Since the ratio $\Delta i_L / \Delta t$ determines the rate of change of current, and the current changes according to the sine law $i_L = I_{\max L} \sin \omega t$, the self-induced emf also follows the same law, with the rate-of-change vector leading the current vector by 90° . Thus, we have

$$e_L = -L\omega I_{\max L} \sin(\omega t + 90^\circ)$$

In order to compensate for the emf, a voltage u_L should be applied to the coil which leads the current by 90° :

$$\begin{aligned} u_L &= -e_L = \\ &= \omega L I_{\max L} \sin(\omega t + 90^\circ) \quad (2-9) \end{aligned}$$

Correspondingly, when a sinusoidal voltage is applied to the coil, a current flows through the coil lagging 90° behind the voltage. When the current reaches its maximum, the rate of its change (as seen from the graphs of the sine function) is zero, thus the voltage also equals zero. As the current decreases, its rate of change increases and, when passing through zero, reaches its maximum. The self-induced emf follows the same law. As a result, the current in the coil circuit is shifted by 90° relative to the voltage, just as in the capacitor circuit, but in the opposite direction (Fig. 2-10). The reason for the current lag in the circuit including an induction coil is that the latter is characterized by certain electric inertia which tends to prevent any current change in the circuit.

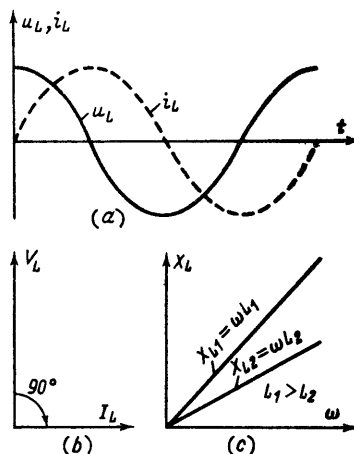


Fig. 2-10. Passing of alternating current through inductance coil (a) voltage and current in inductance coil circuit; (b) vector diagram; (c) inductance coil reactances as functions of frequency

The ratio of the coil voltage to the current through it determines the magnitude of the coil's effective opposition to alternating current. It is also reactive in nature since in an ideal coil there is no conversion of electrical energy into heat, so it is called *inductive reactance*. During a part of a cycle (when the current rises) the coil stores magnetic field energy. This is indicated, among other things, by the fact that the coil is able to pull in magnetic items, thus performing certain work. Then, as the current drops, the coil returns the stored energy back into the circuit (as this takes place the self-induced emf tends to prevent the drop of current in the coil). The inductive reactance is

$$X_L = \frac{V_{\max} L}{I_{\max} L} = \omega L \quad (2-10)$$

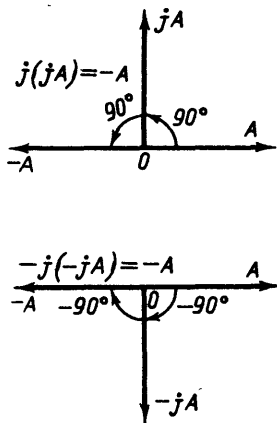


Fig. 2-11. Factor j and its properties

This relationship can be explained as follows. A higher frequency corresponds to a greater rate of change of current and, consequently, gives a greater self-induced emf and inductive reactance. On the other hand, a coil of greater inductance produces a higher self-induced emf and offers a greater reactance to alternating current (see Fig. 2-10).

When alternating current flows through a reactive element, the voltage across the element is shifted in phase relative to the current through it. In vector diagrams, this is usually shown by the angle of 90° between the corresponding vectors. If, as a result of some operation, the vector A rotates by 90° (Fig. 2-11), we shall indicate this turn symbolically with the factor $+j$ (if the vector rotates by 90° counterclockwise) or $-j$ (if the vector rotates by 90° clockwise). If, as a result of the next operation, the vector is again turned by 90° in the same direction, its new direction will be opposite to that of the initial one. The second turn must be also indicated with the multiplier $+j$ or $-j$, hence the resultant vector will be

$$\pm j(\pm jA) = j^2 A = -A \quad (2-11)$$

From Eq. (2-11) the basic property of the factor j follows, that is

$$j^2 = -1 \quad (2-12)$$

Both capacitors and induction coils include conductors which have a definite resistance and, thus, insert losses, i.e., a part of electrical energy turns into heat. Therefore, capacitors and coils are never ideal pure reactances. Also, in capacitors some of electrical energy is dissipated in the dielectric when

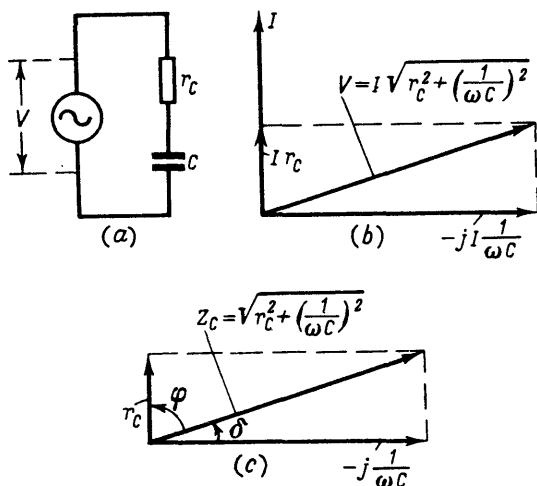


Fig. 2-12. Passing of alternating current through actual capacitor
(a) schematic diagram; (b) vector diagram for voltages and current in the circuit;
(c) impedance diagram

the capacitor is recharged; this loss is caused by re-orientation of molecules which behave similar to dipoles as the electric field direction is changed.

To simplify a vector diagram for a capacitor let us assume that all losses of electrical energy occur in a resistor r_c connected in series with the capacitor C (Fig. 2-12a). The source voltage V is equal to the sum of voltages across the elements of the external circuit. The vector diagram for the moment the current vector is in the vertical position, is shown in Fig. 2-12b. Vector $I r_c$ for the voltage across the resistor is in

phase with the current vector I , and vector $-jI/\omega C$ for the voltage across the capacitor lags by 90° behind the current vector.

The source-voltage vector V is equal to the sum of the two vectors $I r_c$ and $-jI/\omega C$. Since these vectors are at right angles, the magnitude of the vector V is equal to

$$V = \sqrt{(I r_c)^2 + \left(I \frac{1}{\omega C}\right)^2} = I \sqrt{r_c^2 + \left(\frac{1}{\omega C}\right)^2}$$

that is, the hypotenuse of the right-angle triangle.

Dividing all voltages by the common factor I , we obtain an impedance diagram instead of the vector diagram for voltages (Fig. 2-12c). It shows that, when a resistor and a reactance are included into a circuit, the total opposition offered by the circuit to alternating current can be found as a geometrical sum of the corresponding resistive and reactive components displaced relative to each other by 90° . This is called circuit *impedance*.

Specifically, it can be concluded that the resistive and reactive components never cancel each other and that the impedance Z can be zero only if its components are both equal to zero. For the case considered we have

$$Z_c = r_c - j \frac{1}{\omega C} = r_c - j X_c$$

The magnitude or absolute value of the impedance is

$$z_c = \sqrt{r_c^2 + \left(\frac{1}{\omega C}\right)^2} = \sqrt{r_c^2 + X_c^2} \quad (2-13)$$

For a capacitor the ratio of the resistance to the reactance is characterized by

$$\tan \delta = \frac{r_c}{X_c} \quad (2-14)$$

This quantity is called the *loss tangent*; the greater the losses in the capacitor, the greater will be its loss tangent. The value of the loss tangent is given in capacitor specifications. Usually losses in the capacitor are small, thus the loss tangent is also very small.

Similarly, losses in a coil can be presented by an equivalent resistor r_L connected in series with the coil L which has no

losses, i.e. which has only an inductive reactance (Fig. 2-13a). The current flowing through the coil produces a voltage across the inductance, which leads the current vector by 90°

$$V_L = j\omega LI = jIX_L$$

also, it generates a voltage across the resistance, which has the same direction as the current,

$$V_r = Ir_L$$

Vector V for the voltage across the coil leads the current vector I by the angle φ which is less than 90° (Fig. 2-13b).

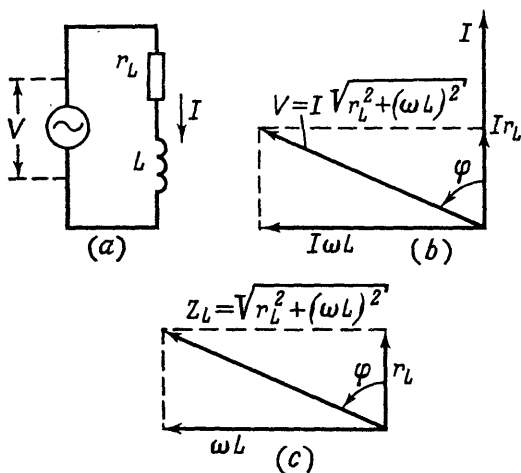


Fig. 2-13. Passing of alternating current through actual inductance
(a) schematic diagram; (b) vector diagram for voltages and current in the circuit;
(c) impedance diagram

Dividing the voltages by the common factor I , we obtain the impedance diagram (Fig. 2-13c). From the diagram, the impedance of the circuit is

$$Z_L = r_L + j\omega L = r_L + jX_L$$

The greater the ratio of the coil reactance to the resistance (which characterizes the losses in the coil), the closer is the angle φ to 90° . This ratio is referred to as *quality factor* or

Q-factor of the coil

$$Q = \frac{X_L}{r_L} \quad (2-15)$$

From the vector diagrams of the circuits including both resistance and reactance (Figs. 2-12 and 2-13) one can see that the tangent of the phase angle between voltage and current in the supply circuit is determined by the ratio of reactance to resistance and the cosine of the same angle is equal to the ratio of resistance to impedance. Thus we have

$$\tan \varphi = \frac{X}{r}; \quad \cos \varphi = \frac{r}{z} \quad (2-16)$$

In what follows, we shall have to deal with complex quantities, so general rules for performing operations on such quantities will be given.

1. Addition of Complex Quantities

Since the vector of a sum of two vectors is equal to the geometrical sum of their projections, it follows that

$$(r_1 + jx_1) + (r_2 + jx_2) = r_1 + r_2 + j(x_1 + x_2)$$

from which it can be seen that to add impedances which are complex quantities, it is necessary to add separately their resistive (real part) and reactive (imaginary part) components. Since complex quantities are represented by vectors two impedances can be equal only if their resistive and reactive components taken separately are also equal.

2. Determination of Reciprocal Complex Quantities

It is often necessary to determine admittance which is the reciprocal of impedance and is defined as

$$Y = \frac{1}{Z} = \frac{1}{r + jx}$$

For this purpose we shall use the factor j and multiply the numerator and the denominator by $r - jx$. Then

$$Y = \frac{r - jx}{(r + jx)(r - jx)} = \frac{r}{r^2 + x^2} - j \frac{x}{r^2 + x^2}$$

It is clear that the admittance is also a complex quantity whose resistive component (real part) is

$$g = \frac{r}{r^2 + x^2} = \frac{r}{z^2} \quad (2-17)$$

and whose reactive component (imaginary part) is

$$b = -\frac{x}{r^2 + x^2} = -\frac{x}{z^2} \quad (2-18)$$

It should be noted that the reactive component of the admittance is opposite in sign to that of the impedance.

3. Products of Complex Quantities

The product of two complex quantities

$$(r_1 + jx_1)(r_2 + jx_2) = r_1r_2 - x_1x_2 + j(x_1r_2 + x_2r_1)$$

is also a complex quantity whose active component is

$$p = r_1r_2 - x_1x_2 \quad (2-19)$$

and reactive component is

$$q = x_1r_2 + x_2r_1 \quad (2-20)$$

2-3. AC POWER

Ac power in an ac circuit is constantly changing. If, however, we divide the cycle of alternating current or voltage into very small time intervals, the values of current and voltage within them can be considered to be relatively unchanged. The ac power dissipated during a short time interval Δt is equal to the product of average values of the current and voltage for this interval

$$\Delta W = iu \Delta t$$

In the general case, the current and voltage in a circuit are phase displaced relative to each other by some angle φ (Fig. 2-14). If we take the moment when the voltage passes through zero to positive values as a time reference, then $v = V_{\max} \sin \omega t$ and $i = I_{\max} \sin(\omega t - \varphi)$.

The ac power dissipated in the circuit during the time interval Δt , is

$$\Delta W = I_{\max} V_{\max} \sin \omega t \sin(\omega t - \varphi) \Delta t$$

Using the following trigonometrical formula $2 \sin \alpha \sin \beta = \cos(\alpha - \beta) - \cos(\alpha + \beta)$, we obtain

$$\Delta W = \frac{V_{\max} I_{\max}}{2} \cos \varphi \Delta t - \frac{V_{\max} I_{\max}}{2} \cos (2\omega t - \varphi) \Delta t$$

The ac power dissipated during the complete cycle of alternating current is a sum of the powers dissipated during all

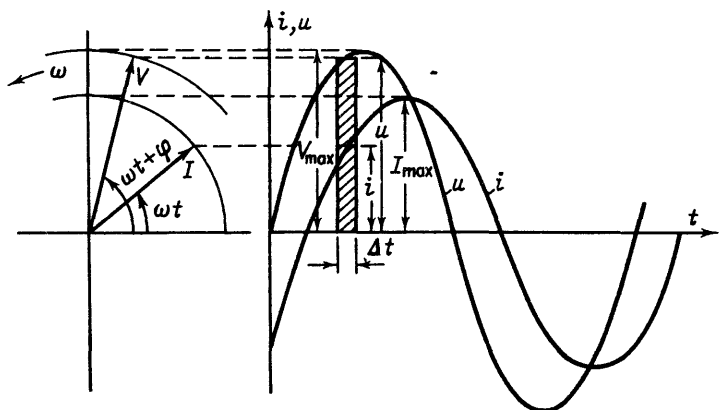


Fig. 2-14. Current, voltage and power for phase angle φ between current and voltage

short time intervals within the cycle *

$$W = \sum \frac{V_{\max} I_{\max}}{2} \cos \varphi \Delta t - \sum \frac{V_{\max} I_{\max}}{2} \cos (2\omega t - \varphi) \Delta t \quad (2-21)$$

Since the first three co-factors of the first term in Eq. (2-21) are constants and the second term after performing the sum-

* Here and elsewhere the Greek letter Σ (sigma) is used to denote the sum of the quantities that follow it. Under this letter one often writes the number (index value) of the first term of the sum, and over it, that of the last term of the sum. For instance,

$$\sum_{i=1}^m a_i = a_1 + a_2 + \dots + a_m$$

If all terms of the sum contain a common factor, it can be placed in front of the summation sign

$$\sum k a_i = k a_1 + k a_2 + \dots + k a_i = k \sum a_i$$

mation over the whole cycle gives zero (the cosine has positive values during one half-cycle, and identical negative values during the other half which cancel each other), it follows that

$$W = \frac{I_{\max} V_{\max}}{2} T \cos \varphi$$

Thus, the average power of alternating current during one cycle is

$$P_{act} = \frac{W}{T} = \frac{1}{2} I_{\max} V_{\max} \cos \varphi \quad (2-22)$$

If the current and voltage are in phase (which occurs when the current passes through resistors), then

$$P_{act} = \frac{1}{2} V_{\max} I_{\max} = \frac{I_{\max}}{\sqrt{2}} \frac{V_{\max}}{\sqrt{2}} \quad (2-23)$$

This expression shows that a power dissipation of the same magnitude will also be given by a direct current and voltage with their values $\sqrt{2}$ times less than the amplitudes of the alternating current and voltage. Therefore, these values are called *effective* values of the alternating current I and voltage V

$$I = \frac{I_{\max}}{\sqrt{2}}; \quad V = \frac{V_{\max}}{\sqrt{2}}$$

For the given amplitudes of current and voltage, the greater the phase angle between them, the less will be the power dissipated. When the phase displacement is 90° (Fig. 2-15) which corresponds to a circuit with reactive elements such as ideal capacitors and lossless induction coils, the average power for one cycle is equal to zero, as these elements are storing energy during one quarter of a cycle returning it back into the source during the next quarter.

However, in this case the concept of a *reactive power* P_{react} developed by a source of an alternating emf during energy exchange with a reactive load is commonly used. This power is defined as

$$P_{react} = \frac{1}{2} I_{\max} V_{\max} \sin \varphi$$

where $I_{\max}$ and $V_{\max}$ are the amplitudes of current and vol-

tage across the load, correspondingly, and φ is the angle between them.

If we take into account that the voltage across a pure reactive load is $V_{\max} = I_{\max} X$, then

$$P_{\text{react}} = \frac{1}{2} I_{\max}^2 X \quad (2-24)$$

or

$$P_{\text{react}} = \frac{1}{2} \frac{V_{\max}^2}{X} \quad (2-25)$$

In radio systems an alternating voltage $u = V_{\max} \sin \omega t$ is often applied to some part of a circuit simultaneously with a direct current and currents of various frequencies which are multiples of ω . In this case we have

$$i = I_0 + I_1 \cos \omega t + I_2 \cos 2\omega t + I_3 \cos 3\omega t + \dots$$

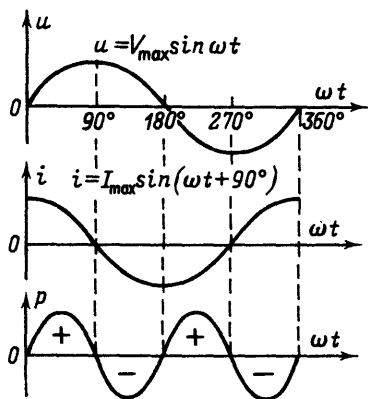


Fig. 2-15. Power as a function of ωt for 90° -phase displacement between current and voltage

What will be the energy effect when these currents will interact with the voltage of the angular frequency ω ? The average power dissipated during one cycle in the result of interaction between the direct current and the alternating voltage will be evidently equal to zero. This occurs due to the

fact that during one half-cycle the source delivers energy to the circuit, and during the other half-cycle the latter returns the same amount of energy to the source. The interaction between the voltage of the frequency ω and the currents of multiple frequencies $n\omega$ is more complex.

Let us determine the average power for the period T . For this purpose, we again divide the whole cycle into very small intervals Δt so that the current and voltage can be considered to be constant within them. For such interval the power is

$$p_i = i(t) u(t)$$

To find the average power for the time period T , we must multiply each p_i by the time interval Δt , add up all the products and divide the result by T . In our case the products to be added will have the form

$$\Sigma V_{\max} I_{\max} \cos \omega t \cos n \omega t \Delta t = V_{\max} I_{\max} \Delta t \Sigma \cos \omega t \cos n \omega t$$

It is easy to prove that all such sums are equal to zero. Fig. 2-16 shows the voltage and the current with the frequency twice as much as that of the voltage ($n=2$) as well as the diagram of products of their instantaneous values. This

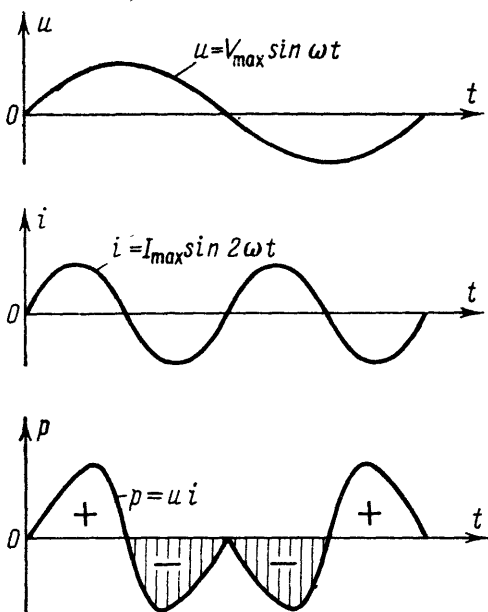


Fig. 2-16. Power as a function of time when current frequency is an integral multiple of voltage frequency

diagram indicates that the instantaneous power also changes periodically and passes from positive values to the identical negative values twice for the time period T . Thus, the average power for the time period T is zero. Clearly, the same will occur with any other combination of multiple frequencies.

On the basis of the above discussion we arrive at the following important conclusion. *If a direct current and alternating currents of multiple frequencies flow in a circuit with an alternating voltage source, then "energy interaction" takes place only with the current whose frequency is equal to that of the source; in the case of dc voltage source "energy interaction" occurs only with the direct component of the current flowing in the circuit.*

2-4. STEADY-STATE AND TRANSIENT PROCESSES IN ELECTRIC CIRCUITS INCLUDING CAPACITORS OR COILS

A charged capacitor possesses a store of potential energy. Now let us find its magnitude. The voltage across the capacitor depends linearly on its charge (Fig. 2-17) since the greater the charge, the higher is the voltage:

$$u = \frac{q}{C}$$

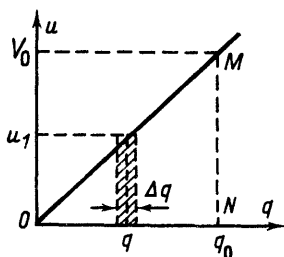


Fig. 2-17. Calculation of electric-field energy stored by capacitor

If a charge q has been accumulated on the capacitor by some moment of time, then its voltage is $u_1 = q/C$. To increase the charge by the value Δq , which is so small that the resultant change in the capacitor voltage can be ignored, certain work must be performed

$$\Delta A = \Delta q u_1$$

This work is numerically equal to the area of the shaded rectangle with sides Δq and u_1 shown in Fig. 2-17.

If the capacitor is further charged up to the voltage V_0 by adding small amounts of charge each equal to Δq , then the total work performed will be

$$A = \Sigma \Delta A = \Sigma \Delta q u$$

This is represented by the area of the triangle OMN which is equal to half the product of its sides

$$A = \frac{1}{2} q_0 V_0$$

or, since $q_0 = CV_0$,

$$A = \frac{1}{2} CV_0^2 \quad (2-26)$$

Magnetic-field energy accumulated by a coil depends on the value of the current flowing through the coil. The law which is followed by the current in the coil reaching up to its steady-state value, is here of no importance. Let us assume that the current increases according to a linear law (Fig. 2-18). In this case the work to be performed during the short time interval Δt when the charge Δq is passing through the coil

$$\Delta A = e_L \Delta q$$

where $\Delta q = i_1 \Delta t$, and i_1 is the average current value for the interval Δt . Using expression (2-8) we can write

$$\Delta A = -L \frac{\Delta i}{\Delta t} i_1 \Delta t$$

To increase the current in the coil up to the value I_0 some work must be performed, which is equal to the sum of those performed in the separate time intervals

$$A = -\Sigma L i_1 \Delta t \frac{\Delta i}{\Delta t}$$

Here the minus sign indicates that some external source, which feeds the coil, loses its energy as the current in the coil increases. In the above sum, the constant factors L and $\Delta i / \Delta t$ can be placed in front of the summation sign. The latter factor is constant because the rate of change of the current remains, by assumption, unchanged, i.e. $\Delta i / \Delta t = I_0 / t_0 = \text{const.}$ Hence,

$$A = -L \frac{I_0}{t_0} \Sigma i_1 \Delta t$$

As seen from Fig. 2-18, the product $i_1 \Delta t$ is determined by the area of the shaded rectangle and $\Sigma i_1 \Delta t$, by the area of the

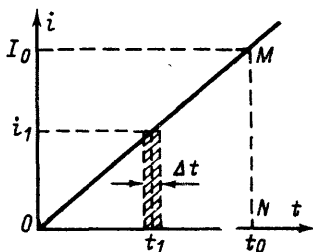


Fig. 2-18. Calculation of magnetic-field energy stored by coil

triangle OMN , which is equal to half the base times the height of the triangle,

$$\Sigma i_1 \Delta t = \frac{I_0 t_0}{2}$$

Accordingly, the work to be performed by the source equals

$$A = -L \frac{I_0}{t_0} \frac{I_0 t_0}{2} = -L \frac{I_0^2}{2} \quad (2-27)$$

This work is expended in energy accumulated by the coil's magnetic field.

Consider now a circuit including a capacitor. When the voltage of source which feeds the circuit is changed, the voltage across the capacitor cannot change instantaneously as this would mean that the energy stored in the capacitor is changed by a finite value for an infinitely small period of time, in other words, the source would produce infinite power, which is impossible. Consequently, any capacitor always charges and discharges gradually, and its voltage reaches a new steady-state value only after a time lag. It should be noted that the current in a capacitive circuit can change abruptly as it does not determine the amount of energy stored in the capacitor.

The process of transition from one steady state to another is called *transient* or *non-stationary process*. It arises in circuits which include energy storing elements when operating conditions in the circuit or supply are changed. Among these are circuits containing induction coils. Thus currents and voltages can change almost instantaneously only in circuits which consist of pure resistors.

A transient process continues during a finite interval of time, while the system changes over from one steady state to another with the latter being determined by new operating conditions. The duration of transients in a given electric system characterizes its time delay, in other words, its ability to change over more or less quickly from one steady-state condition to another. Similar phenomena are also known from physics. For instance, any real mechanical system does not stop instantly but reduces its speed gradually down to zero after an external tractive force applied to it has ceased its action. When a body is heated, its temperature never rises at once but slowly increases to a certain value. After transients have

completed, the system reaches the steady-state condition corresponding to the new operating conditions, which it maintains until external conditions are again changed.

Let us examine the process of capacitor charge and discharge taking into account the circuit's total resistance (which includes resistance of connecting wires, internal resistance of the supply or of a specially connected resistor). At the moment t_0 the switch K is closed and the capacitor charge current flows in the circuit (Fig. 2.19). At first the capacitor

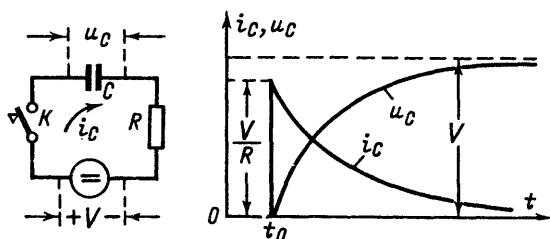


Fig. 2-19. Current and voltage in capacitive circuit; capacitor is charged from a dc source through a resistor

plates have no charges, therefore nothing prevents the charging, thus the current has its maximum value which is only limited by the resistance R of the circuit. As the electric charges begin to accumulate on the plates they tend to prevent further flow of charges so that the current decreases and ceases completely when the voltage across the plates becomes equal to and opposing that of the supply.

The less the charge current, i.e. the higher the circuit resistance, the longer will the voltage across the capacitor reach its steady-state value. It is also evident that the higher the capacitance of the capacitor, the longer the charging process will continue because a greater charge must be accumulated to raise the voltage across the capacitor up to that of the supply. Therefore, the duration of this transient process is proportional to the product

$$RC = \tau \quad (2-28)$$

which is called *time constant* of the circuit. It is easy to show that it has the dimension of time

$$[RC] = \frac{[V][q]}{[I][V]} = [t]$$

In a time of $t=\tau$ seconds (from the beginning of the capacitor charging process) the voltage across the capacitor reaches 63% of the supply voltage; in a time of $t=2\tau$ seconds it rises up to 87%; and in a time of $t=3\tau$ seconds it approaches 95%. Thus, the transient process is essentially completed in the time of 3τ .

The rate of voltage rise and current fall in the circuit follows the so-called *exponential law*, illustrated in Fig. 2-19,

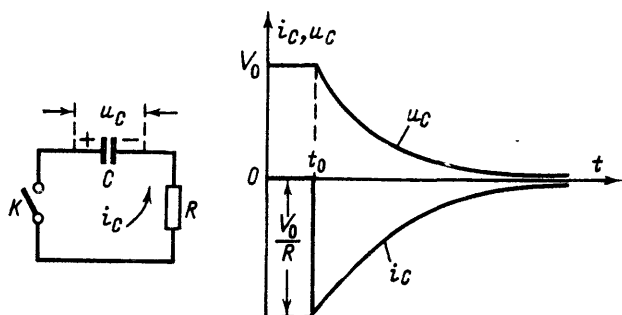


Fig. 2-20. Current and voltage in capacitive circuit; capacitor is discharged through a resistor

having a rapid initial change but approaching final values quite slowly. This law expressed mathematically is

$$\left. \begin{aligned} i &= I_0 e^{-\frac{t}{\tau}} \\ u_C &= V_0 \left(1 - e^{-\frac{t}{\tau}} \right) \end{aligned} \right\} \quad (2-29)$$

where $e=2.7$ is the base of natural logarithms, and $I_0=V_0/R$. This law is valid for the vast majority of real transient processes, the only difference usually being the value of the time constant.

If a capacitor C charged to the voltage V_0 is discharged through resistor R (Fig. 2-20) the discharge current at the initial moment t_0 is maximum, and then both voltage and current fall as the discharge progresses. The less the value of the discharge current, i.e. the higher the circuit resistance and the greater the charge stored in the capacitor (i.e. the higher its capacitance), the longer will last the discharge. Corres-

pondingly, the duration of the discharge process will be proportional to the product $R \cdot C$. It follows that the time constant does not depend on the character of the process, i.e. whether it results in storing or expenditure of energy, but is solely determined by the component values of the circuit.

Induction coils also accumulate energy which is concentrated in their magnetic field whose intensity is proportional to the current flowing in the coil. From this it can be inferred that the current in an inductive circuit cannot vary instantaneously and will change gradually provided conditions for this change exist. In other words, the transition to the new steady state requires a finite time.

If a coil is connected to the source of the dc voltage V via the resistor R (Fig. 2-21), the current through the coil, in accordance with the general rule stated above, will be zero at the instant the circuit is closed. This is possible provided the

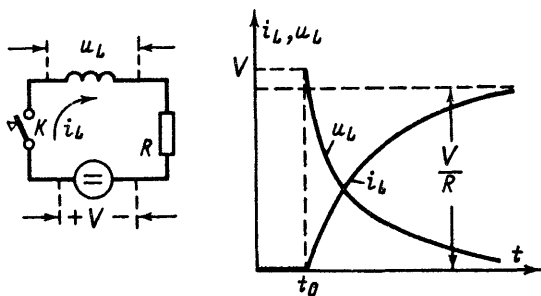


Fig. 2-21. Current and voltage in inductive circuit; coil is connected to a dc source through a resistor

self-induced emf at this moment is equal and opposite to the source voltage. Gradually, the current will rise approaching its steady-state value

$$I_{st} = \frac{V}{R} \quad (2-30)$$

On the other hand, the self-induced emf will decrease and become zero by the end of the transient process when direct current starts to flow through the coil.

Obviously, the transient process will be the longer, the greater the amount of energy to be accumulated by the coil in

the steady-state condition. The stored energy, in its turn, will be the greater, the higher the inductance and the greater the current flowing through the coil in the steady-state conditions which is inversely proportional to the resistance of the

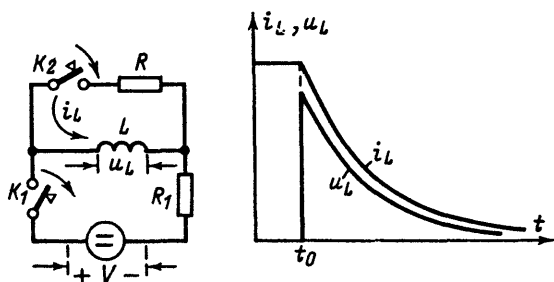


Fig. 2-22. Current and voltage in inductive circuit; coil is connected to a resistor

circuit. Thus, for a given inductive circuit the time constant

$$\tau = \frac{L}{R} \quad (2-31)$$

It is easy to show that this ratio also has the dimension of time

$$[\tau] = \frac{[L]}{[R]} = \frac{\left[\frac{V}{I}\right] [t]}{\left[\frac{V}{I}\right]} = [t]$$

Let us now examine the reverse process which takes place when a coil carrying a current is connected to a resistor. Let us assume that the coil was at first connected to a supply and some direct current was flowing in it (Fig. 2-22). At the instant the switch K_1 opens, the switch K_2 closes and the current in the circuit begins to decrease. As a result, a very large emf is induced which tends to prevent the current decrease so that the magnetic field energy stored by the coil is transformed into energy of the electric current flowing in the circuit through the resistor R .

It is clear that the greater the magnetic field energy stored by the coil, i.e., the higher its inductance, the longer will be

the duration of the transient process. Also, the higher the resistance of the discharge circuit, the less will be the rate of change (in our case, of the current decrease). Therefore, the time constant of the circuit in this case will also be $\tau=L/R$.

2-5. NONSINUSOIDAL CURRENTS AND THEIR SPECTRA

We described above the peculiarities of sinusoidal currents flowing through the basic components of electric circuits, such as wires, coils and capacitors. Sinusoidal currents, however, are only one particular form of alternating currents used in radio systems. Some typical forms of alternating currents encountered in radio circuits are shown in Fig. 2-23. At first sight, these seem to raise insuperable obstacles in the way of studying the phenomena occurring in radio circuits because each of these currents behaves in the same circuit components in a different manner. What shall we do in such cases?

Mathematicians suggested the answer. They proved that any periodic current having a period T can be presented by an infinite sum of a direct and sinusoidal (harmonic) currents with different amplitudes, frequencies and initial phases

$$i(t) = I_0 + I_{\max 1} \cos\left(\frac{2\pi t}{T} + \varphi_1\right) + I_{\max 2} \cos\left(2\frac{2\pi t}{T} + \varphi_2\right) + \\ + I_{\max 3} \cos\left(3\frac{2\pi t}{T} + \varphi_3\right) + \dots \quad (2-32)$$

Here I_0 is the dc component of the current and is equal to its average value taken over the period T . For the current in the form of alternating pulses with equal amplitudes but opposite signs, illustrated in Fig. 2-23a, we have $I_0=0$.

Note that the sinusoidal currents in expression (2-32) have frequencies that are integral multiples of the lowermost angular frequency $\omega=2\pi/T$. The sinusoidal component with this frequency is referred to as *first harmonic* or *fundamental component*; the component with the double frequency $2\omega=2\times 2\pi/T$, as *second harmonic*, etc. All harmonics beginning with the second are called *higher harmonics*. Of course, one may ask if the representation of a nonsinusoidal current in the form of a sum of harmonic currents will really facilitate the solution of the problem in view of the fact that the sum itself

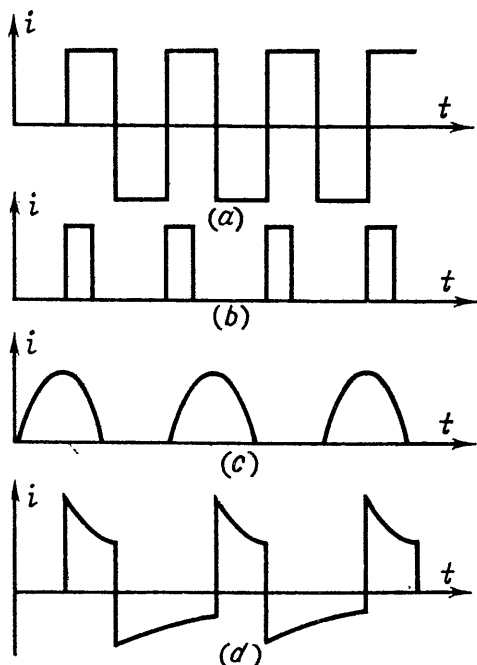


Fig. 2-23. Typical waveforms of periodic nonsinusoidal currents
 (a), (b) and (c) symmetrical waveforms; (d) asymmetrical waveform

is infinite. Mathematicians, however, who found the simple way of calculating the amplitudes of harmonic components have shown that they diminish (though not always monotonically) with an increase of the harmonic number (Fig. 2-24). Therefore, in most cases only a finite number of harmonics is required. This greatly simplifies the problem as the properties of direct and sinusoidal currents are well known.

Yet another question arises: how close do these mathematical concepts approximate physical realities and can they be used in engineering practice? To answer this, we make the following experiment. We connect a tunable oscillatory circuit (OC) with a meter I_{OC} to a source of nonsinusoidal periodic current (NCS) (Fig. 2-25a). When the oscillatory circuit is tuned over a range of frequencies, current appears in it only at

the frequencies ω , 2ω , 3ω , etc., as foretold by mathematicians, and the amplitudes of these currents decrease with an increase of frequency taking certain precalculated values (Fig. 2-25b).

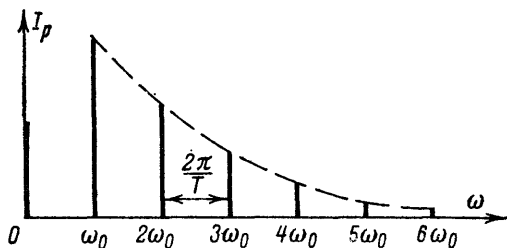


Fig. 2-24. Spectrum of periodic nonsinusoidal current

The experiment can be somewhat improved to make sure that the harmonics exist simultaneously by connecting a dc meter I_0 and a set of circuits tuned to frequencies ω , 2ω , 3ω , etc. into the circuit of *NCS* (Fig. 2-25c). The result will be the same.

These experiments seem to prove that harmonics do actually exist. However, this is not true. The above statement is similar to one witty saying that there exists a Venus of Milo in every block of marble. It would be more correct to say that there exists a single nonsinusoidal current which, however, excites different circuits as if it consisted of an infinite number of harmonic currents having certain amplitudes, frequencies and phases. Therefore, from a practical viewpoint we can assume that harmonics do really exist and use this concept in this or that way, which will never come into contradiction with practice.

On the basis of the above considerations a conclusion can be drawn that any nonsinusoidal periodic current consists of a set of sinusoidal currents having multiple frequencies, which is usually referred to as *spectrum*. The actual width and composition of the spectrum depend on the period of the current and its waveform. For example, if some reference point can be found on a nonsinusoidal current graph relative to which all current values on the right are equal to those at equidistant points on the left (such waveforms are called symmetrical;

some examples of these are shown in Figs. 2-23 *a*, *b* and *c*), then all initial phases in expression (2-32) will be equal to 0° or 180° , thus it can be rewritten as

$$i(t) = I_0 \pm I_{\max 1} \cos \omega t \pm I_{\max 2} \cos 2\omega t \pm I_{\max 3} \cos 3\omega t \pm \dots \quad (2-33)$$

The closer the current waveform to a sinusoid, the fewer the harmonics needed for summation to approximate satisfactorily the exact waveform of the nonsinusoidal current. To

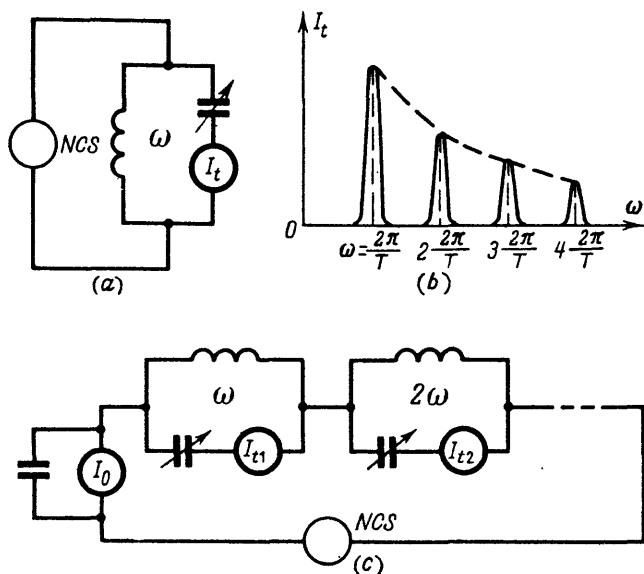


Fig. 2-25. Spectrum analysis of nonsinusoidal periodic current

reproduce graphically signals which have sharp waveforms harmonic components with very high frequencies, i.e. harmonics of very high order, should be used since harmonic waveforms are rather smooth.

To illustrate the above we add the components described by the first three terms in expression (2-33), first for the case of pulses in the form of a half-sinusoid (Fig. 2-26) and then for square pulses (Fig. 2-27). It is clear that in the first case

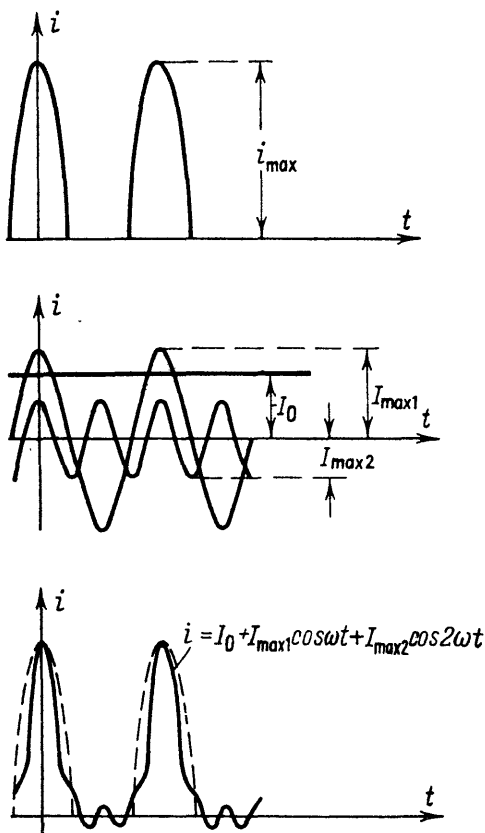


Fig. 2-26. Approximation of half-sinusoid pulses by the first three components

the summation will give a far better approximation to the real current waveform than in the second. To approximate satisfactorily square pulses, a much greater number of harmonic components is required. Thus, the actual spectrum of square pulses proves to be much wider than that of half-sinusoid pulses.

As concerns the width of the actual spectrum, a precise answer is given by a well-developed mathematical method known as

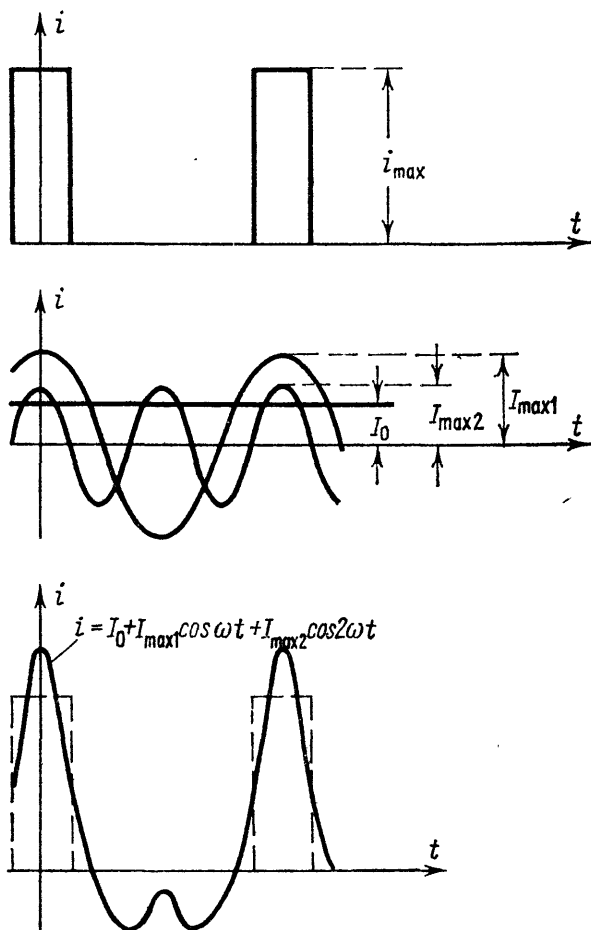


Fig. 2-27. Approximation of square pulses by the first three components

harmonic analysis. It allows the amplitudes of higher harmonics, which decrease with increasing frequency as well as the harmonic's phases to be evaluated. This has been done for the majority of signal waveforms encountered in radio circuits and the results are usually presented in the form of amplitude-

frequency (Fig. 2-28a) and phase-frequency (Fig. 2-28b) spectra that can be found in radio engineering handbooks.

The spectrum width of a pulse signal can roughly be estimated from its duration. In fact the pulse waveform is essentially determined by the harmonic whose half-cycle is equal to the pulse duration (Fig. 2-29), i.e. $T = 2\tau$, where τ is the pulse duration. To approximate satisfactorily this pulse waveform yet another two or four har-

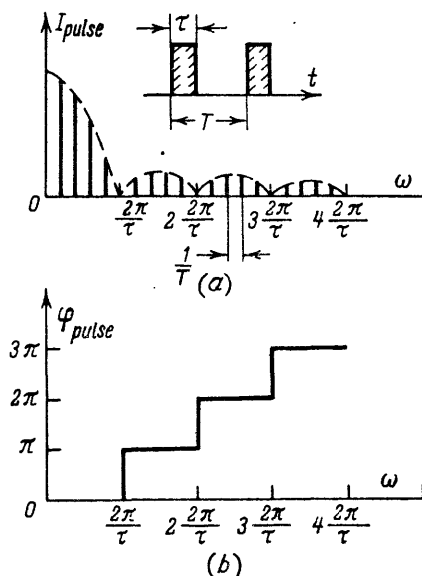


Fig. 2-28. Spectra of a series of square pulses

(a) amplitude-frequency spectrum; (b) phase-frequency spectrum

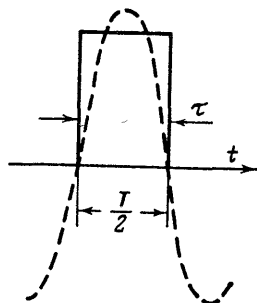


Fig. 2-29. Square pulse and sinusoid approximating it

monics must be added. Thus, the highest harmonics which must be included in the spectrum of the pulse signal and which determine its spectrum width will have the frequency

$$f = \frac{2.4}{T} = \frac{1.2}{\tau} \quad (2-34)$$

The above method provides answers to all questions concerning nonsinusoidal periodic currents in radio circuits. Besides these, however, unwanted signals also exist such as single nonperiodic currents generated by lightning, pulses appearing during circuit switching, etc. How can we analyze their in-

fluence on the performance of radio equipment? The harmonic-analysis method described above can be extended even to such signals by considering a single nonperiodic signal as a periodic one with period T approaching infinity. What will happen in this case to the spectrum of such a signal? It is easy to make sure that as the period increases, the frequency

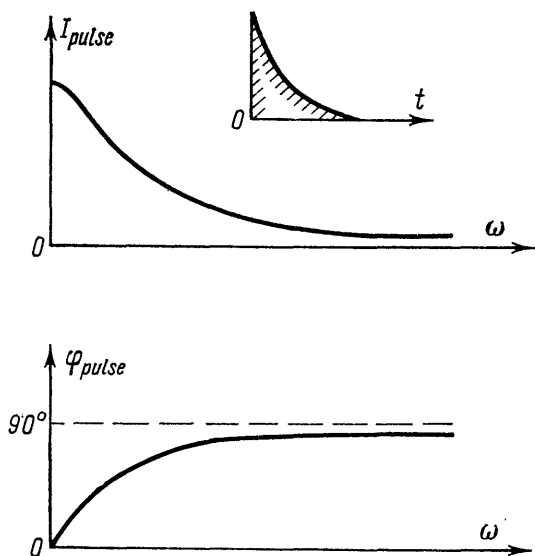


Fig. 2-30. Single pulse and its spectra

difference between adjoint harmonics lessens. In fact, if we take the n th and $(n+1)$ th harmonics, the difference between their frequencies

$$\Delta\omega = (n+1)\omega - n\omega = \omega = \frac{2\pi}{T}$$

tends to 0 as $T \rightarrow \infty$.

Consequently, the line spectrum of periodic signals turns into the continuous spectrum of a single (nonperiodic) signal. This explains, for example, the fact that an atmospheric discharge can be heard in radio receivers tuned to quite different frequencies. However, the amplitudes of the spect-

rum components, just as in the case of periodic signals, decrease with increasing frequency. Therefore, in the ultrashort-wave band the components of atmospheric discharges have already such a low intensity that they hardly hinder reception. This is one of the reasons why local radio stations broadcast in this band. Fig. 2-30 shows a typical signal and its amplitude-(above) and phase-frequency (below) spectra.

Thus, the harmonic-analysis method enables us to reduce the problem of any signals passing through radio circuits to that of a series of sinusoidal signals. Therefore, in the subsequent sections emphasis is placed on harmonic voltages acting on individual circuits and devices. This will make it possible to understand the detailed behaviour of other, more complicated signals in various devices provided the signal spectra are known.

2-6. FREE OSCILLATIONS IN A CIRCUIT

Let us examine the processes that take place in a circuit consisting of a capacitor and an induction coil (Fig. 2-31) which are both lossless. Open switch K , close switch K_1 and charge the capacitor from a dc source. Then open switch K_1 and close switch K . The capacitor at once begins to discharge through the coil. As previously in the coil there has been no current, the current i will gradually increase from zero due to the action of the self-induced emf, whereas the voltage u_C across the capacitor will decrease. Hence, the electric field energy stored in the capacitor lowers and that of the coil magnetic field increases.

When the voltage across the capacitor falls to zero, all the energy of the capacitor electric field has been transformed into the energy of the coil magnetic field, which reaches its maximum. Simultaneously, the current flowing in the coil also reaches its maximum value. Since the current in the coil cannot suddenly drop to zero, it will decrease gradually in the same direction, recharging the capacitor, i.e. feeding its plates with charges of the opposite sign relative to that they have had before. When the current falls to zero, all the energy of the coil magnetic field has transformed into that of the capacitor electric field, and the voltage across the capacitor takes the same value as at the very beginning of the process, but

with the opposite sign. Then the process is repeated with the discharge progressing in the opposite direction.

In such a way, sustained oscillations will occur in the circuit as the energy of the capacitor electric field is transformed into that of the coil magnetic field, and vice versa. Therefore, such a circuit consisting of a capacitor and an induction coil is called *oscillatory circuit*. As there are no losses in an ideal

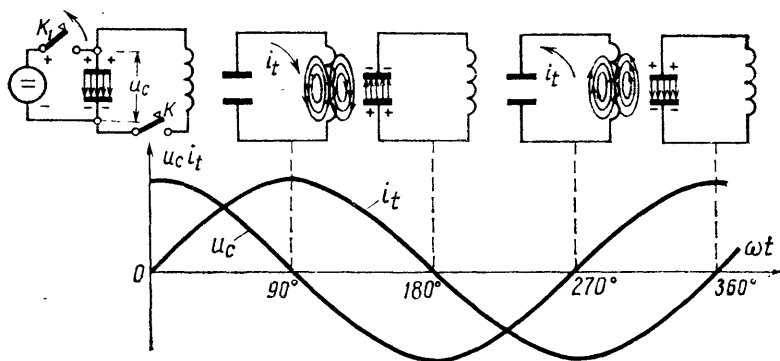


Fig. 2-31. Free oscillations in a circuit

circuit, this process continues infinitely. Analogous phenomena can be observed in mechanical oscillatory systems such as an ideal pendulum where potential energy at the upper position is transformed into kinetic energy at the lower position, and vice versa.

The process described bears the name of *free oscillations* since it proceeds in the circuit without the influence of any external exciting force but solely due to some amount of energy stored in one of the circuit elements. The current and voltage across the circuit, as can be seen from Fig. 2-31, vary sinusoidally but with a phase displacement of 90° .

Since the maximum energies stored in the capacitor and coil are equal to each other, $V_{\max C}^2 C/2 = I_{\max L}^2 L/2$, where $V_{\max C}$ is the voltage across the capacitor, $I_{\max L}$ is the current in the coil, C is the capacitance and L is the inductance, we have

$$\frac{V_{\max C}}{I_{\max L}} = \sqrt{\frac{L}{C}} = \rho \quad (2-35)$$

This quantity is referred to as *circuit effective resistance*; it represents the effective resistance of the coil or capacitor to the current flowing through them (since the voltages across both elements are equal and the same current flows through them, their effective resistances are also equal), and is defined as

$$\rho = \omega_0 L = \frac{1}{\omega_0 C} \quad (2-36)$$

Thus, when free oscillations are maintained in the circuit the effective resistances of both the capacitor and coil are equal so that a current of one particular frequency can flow in the circuit; for this frequency

$$\omega_0 L = \frac{1}{\omega_0 C} \quad (2-37)$$

where ω_0 is the angular frequency of free oscillations also called *natural* oscillations. From (2-37) the following expression for angular frequency of free oscillations is obtained

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad (2-38)$$

Here ω_0 is in radians per second, L in henries, and C in farads.

The frequency of circuit free oscillations (in hertz) can be found from the formula

$$f_0 = \frac{1}{2\pi\sqrt{LC}} \quad (2-39)$$

and the period of free oscillations is

$$T = \frac{1}{f_0} = 2\pi\sqrt{LC} \quad (2-40)$$

From (2-40) it follows that the free period T increases with the increase of the inductance and capacitance in the circuit. This occurs due to the fact that the higher the inductance, the slower changes the current and the higher the capacitance, the greater the time required for recharging the capacitor.

So far we have considered oscillations in a circuit without any losses. In real circuits, however, energy is always lost due to heating of wires and dielectrics, as well as through radiation resulting in a gradual decrease of the amplitude of oscillations.

lations. Such oscillations illustrated in Fig. 2-32 are called damped oscillations. When designing a circuit, it is assumed that it includes a lumped resistor r_t which dissipates the same amount of power as is consumed in the circuit due to all kinds of losses.

The quicker the oscillation amplitude decrease, the greater the amount of the total energy originally stored in the circuit lost during the oscillation period, i.e. the greater is the lumped resistance as compared with the circuit's effective resistance. The ratio of the circuit's effective resistance ρ to its lumped resistance r_t

$$Q = \frac{\rho}{r_t} = \frac{\omega_0 L}{r_t} = \frac{1}{\omega_0 C r_t} \quad (2-41)$$

is called quality factor (Q -factor) of the circuit.

Since the dissipated power is proportional to the lumped (ohmic) resistance of the circuit, and the reactive power provided by its reactive elements is proportional to their reactances, the quality factor Q is also a measure of the ratio of these two powers. For instance, if two circuits have the same free-oscillation frequency but different Q 's, then free oscillations in the circuit with the higher Q will be damped slower than in the other. Circuits used in radio engineering usually have quality factors ranging from 50 to 300.

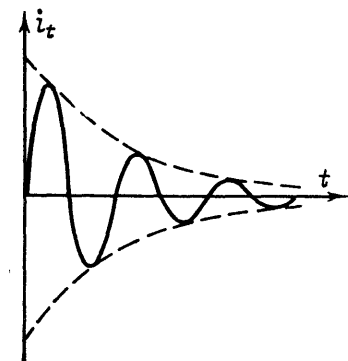


Fig. 2-32. Damped oscillations

2-7. FORCED OSCILLATIONS IN A SERIES-CONNECTED CIRCUIT

When a source of alternating voltage is connected in series with a circuit consisting of a coil and a capacitor (Fig. 2-33a), a current flowing in the circuit will have the frequency of the source, which generally differs from that of the circuit free oscillations. In this case *forced*, rather than free, *oscillations* occur in the circuit.

Figure 2-33*b* presents a vector diagram characterizing the forced-oscillation mode of the circuit. When plotting this diagram, one should begin with the vector of the current I_t which passes successively through all the circuit elements. Let us assume that at a certain moment this vector is in the horizontal position and directed to the right. The voltage V_r

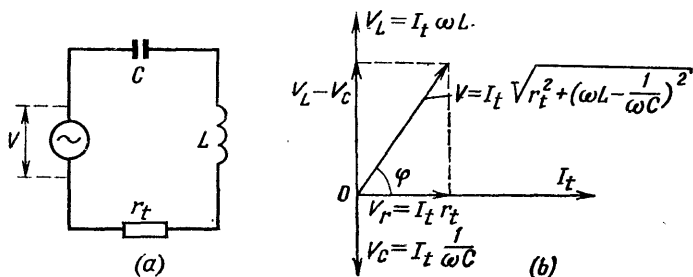


Fig. 2-33. Series connection of an alternating emf source into a circuit
(a) schematic diagram; (b) vector diagram

across the resistor r_t (which characterizes the total losses in the circuit) is in phase with the current I_t and equal to

$$V_r = I_t \cdot r_t$$

Vector V_L of the voltage across the coil leads the current vector by 90° and is equal to

$$V_L = I_t X_L = I_t \omega L$$

Vector V_C of the voltage across the capacitor lags behind the current vector by 90° and is equal to

$$V_C = I_t \frac{1}{\omega C}$$

Thus, the voltages across the coil and the capacitor are displaced by 180° relative to each other, in other words, are in antiphase. Therefore, they, to some degree, compensate each other. For this reason the vectors V_L and V_C are shown in the vector diagram in opposite directions. The resultant vector has the direction of the larger of the two vectors, and its magnitude is equal to the difference between them. If the vector is directed to the right and upwards, it is assumed to be positive,

and the voltage across the capacitor is considered to be negative. Thus, we have

$$V_L - V_C = I_t \omega L - I_t \frac{1}{\omega C} = I_t \left(\omega L - \frac{1}{\omega C} \right)$$

To determine the total voltage V across the circuit the voltage V_r across the resistor r_t must be added to the voltage $V_L - V_C$. As they are displaced by 90° relative to each other, it follows that

$$V = \sqrt{V_r^2 + (V_L - V_C)^2} = I_t \sqrt{r_t^2 + \left(\omega L - \frac{1}{\omega C} \right)^2}$$

From this expression the modulus of the circuit impedance is defined as

$$z_t = \frac{V}{I_t} = \sqrt{r_t^2 + \left(\omega L - \frac{1}{\omega C} \right)^2} \quad (2-42)$$

As can be seen from the vector diagram, the total current I_t in the circuit is shifted relative to the source voltage by the angle φ which can be found from the formula

$$\tan \varphi_t = \frac{V_L - V_C}{V_r} = \frac{X_t}{r_t} \quad (2-43)$$

where

$$X_t = \omega L - \frac{1}{\omega C} \quad (2-44)$$

The greater the angle φ_t , the greater the total reactance X_t of the circuit and the smaller its total resistance r_t .

These expressions show that the reactance ωL of the coil, the reactance $1/\omega C$ of the capacitor as well as the circuit's total reactance $\omega L - 1/\omega C$ and its behaviour (i.e., whether it is inductive or capacitive) — all depend on the frequency of the source (Fig. 2-34).

It is seen from Fig. 2-34 that at the frequency corresponding to the equal inductive and capacitive reactances the circuit total reactance is zero. Since in this case $\omega L = 1/\omega C$, the ac source angular frequency ω is equal to $1/\sqrt{LC}$, i.e. to the frequency of the circuit free oscillations described by (2-38). Also, the circuit inductive and capacitive reactances are equal to the circuit effective resistance ρ , just as in the case of free oscillations.

Thus, at this frequency the circuit has only a pure resistance which is equal to the total resistance r_t , i.e. its reactances cancel each other. This circuit condition is called *resonance*, and the frequency at which this takes place is termed as resonance frequency. As the resistance r_t is usually very small, the current $I_{t,r}$ in the circuit at resonance is high and equal to

$$I_{t,r} = \frac{V}{r_t} \quad (2-45)$$

This current will be the higher, the smaller the circuit's total resistance r_t .

As the current in the circuit increases, voltages across the coil and the capacitor are also increased; whereas their sum remains zero. Thus, we have

$$V_{L,r} = -V_{C,r} = I_{t,r} \cdot \rho = V \frac{\rho}{r_t} = VQ \quad (2-46)$$

Hence, the circuit quality factor Q also indicates how many times the voltage across the capacitor (or coil) at resonance is greater than the source voltage.

In the case considered, the current source is connected in series with the circuit, and, at resonance, voltages across the reactive elements of the circuit sharply increase which is very important. Therefore, the resonance condition of such a circuit is called *series* or *voltage resonance*.

At frequencies below the resonance frequency, the capacitive reactance, which increases with decreasing frequency, becomes dominant. For this reason the circuit behaves like an equivalent capacitor and resistor connected in series, the capacitance of the equivalent capacitor differing from that of the circuit capacitor.

As the operating frequency deviates farther from the resonance one, the absolute value of the difference $\omega L - 1/\omega C$

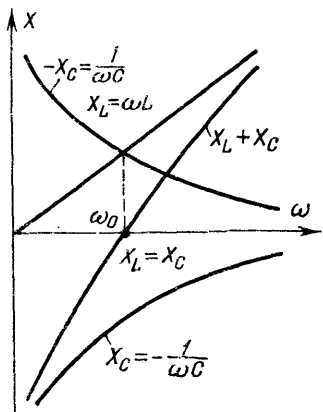


Fig. 2-34. Reactances of circuit components as functions of frequency

rapidly increases, hence the circuit impedance also rises very quickly, while the circuit current decreases. This deviation also causes an increase in the phase shift of the circuit current with respect to the source voltage, and when the difference between the two frequencies becomes large enough the circuit starts to behave essentially like a capacitor with a high reactance. The difference between the source and resonance frequencies is called *detuning increment* and denoted as

$$\Delta\omega = \omega - \omega_0 \quad (2-47)$$

On the other hand, at frequencies higher than the resonance frequency the reactance of the coil exceeds that of the capacitor so that the circuit becomes equivalent to a combination of a resistor and a coil connected in series, the inductance of the equivalent coil differing from that of the circuit coil.

As the frequency increases and the detuning increment $\Delta\omega$ becomes larger, the circuit impedance rapidly grows and the circuit current falls. Also, the phase shift of the circuit current with respect to the source voltage increases, and when $\Delta\omega$ becomes large enough the circuit impedance has essentially the inductive character.

For a fixed amplitude of the voltage V , the ratio of the current in the detuned circuit to the current at resonance is inversely proportional to the ratio of the detuned circuit impedance to that at resonance. Thus, we can write

$$\frac{I_t}{I_{t,r}} = \frac{z_r}{z} = \frac{r_t}{\sqrt{r_t^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \quad (2-48)$$

Rearranging the expression under the root sign

$$\omega L - \frac{1}{\omega C} = \omega L \left(1 - \frac{1}{\omega^2 LC}\right)$$

and taking into account that $LC = 1/\omega_0^2$, we have

$$\omega L - \frac{1}{\omega C} = \omega L \frac{\omega^2 - \omega_0^2}{\omega^2} = \omega L \frac{(\omega - \omega_0)(\omega + \omega_0)}{\omega^2}$$

At frequencies close to the resonance frequency, where it can be assumed that $\omega \approx \omega_0$ and $\omega_0 + \omega \approx 2\omega_0$, we obtain by using formula (2-47):

$$\omega L - \frac{1}{\omega C} \approx \frac{2\Delta\omega}{\omega_0} \omega_0 L \quad (2-49)$$

Substituting this expression into (2-48) and dividing the numerator and denominator by r_t we have

$$\frac{I_t}{I_{t,r}} = \frac{z_r}{z} = \frac{1}{\sqrt{1 + \left(\frac{\omega_0 L}{r_t}\right)^2 \left(\frac{2\Delta\omega}{\omega_0}\right)^2}}$$

Taking into account that $\omega_0 L/r_t = Q$, this expression can be rewritten as

$$\frac{I_t}{I_{t,r}} = \frac{z_r}{z} = \frac{1}{\sqrt{1 + Q^2 \left(\frac{2\Delta\omega}{\omega_0}\right)^2}} \quad (2-50)$$

Expression (2-50) describes the resonance characteristic (curve) of the circuit. It shows that for the same detuning increment the relative drop of current in the circuit will be the greater, the higher the circuit quality factor. If we use a circuit with a sufficiently high Q (of the order of several dozens or more, which is always practicable), the factor Q^2 under the root sign in the expression (2-50) becomes very large, and even for a very small $\Delta\omega$ the circuit current will drop sharply as compared to its value at resonance. This is illustrated in Fig. 2-35, where resonance curves for circuits with different quality factors are shown. They indicate that high amplitude oscillations in the circuit can be excited only by sources whose frequency is close to that of the circuit free oscillations. In other words, the circuit transmits a certain band of frequencies. This property is characterized by the *circuit passband*, i.e. by the band of frequencies within which the oscillation amplitude is decreased less than $\sqrt{2}$ times as compared with that at resonance.

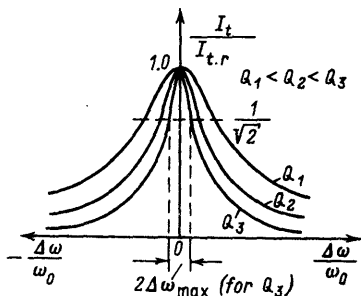


Fig. 2-35. Resonance curves for circuits with different Q -factors

The maximum detuning increment $\Delta\omega_{\max}$ within the circuit passband can be found from (2-50):

$$\frac{I_t}{I_{t,r}} = \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{1 + Q^2 \left(\frac{2\Delta\omega_{\max}}{\omega_0} \right)^2}},$$

hence the ratio of $\Delta\omega_{\max}$ to the resonance frequency ω_0 of the circuit is

$$\frac{\Delta\omega_{\max}}{\omega_0} = \frac{1}{2Q}$$

As the resonance curve is symmetrical, the circuit passband has a width equal to

$$2\Delta\omega_{\max} = \frac{\omega_0}{Q} \quad (2-51)$$

Thus, the higher the quality factor of the circuit, the narrower its passband.

The quality factor is mainly determined by the inductance coil quality which accounts for almost all losses. The bandwidth of the circuit $2\Delta\omega_{\max}$ is 0.3 to 2 per cent of the circuit resonance frequency. The ability of the circuit to transmit currents with frequencies close to the resonance one and to reject currents of other frequencies is called *selectivity*. The narrower the circuit passband, i.e. the higher the quality factor, the greater the circuit selectivity.

Selective properties of tuned circuits are widely used in radio engineering, for instance, when it is necessary to separate a desired signal frequency to which a particular receiver is tuned from all other signals transmitted by various radio stations and fed to the input of the receiver via its antenna.

2-8. FORCED OSCILLATIONS

IN A PARALLEL-CONNECTED CIRCUIT

In radio devices, oscillatory circuits are often employed in which elements (capacitor and coil) are connected in parallel to an emf source (Fig. 2-36). Such circuits are called *parallel circuits*.

When plotting a vector diagram for a parallel circuit, vector V of the voltage across the circuit is plotted along the ho-

horizontal axis (Fig. 2-37). With voltage applied, the current I_C flows through the capacitor, leading the voltage approximately by 90° (since losses in a capacitor are usually very small and not considered). This current is

$$I_C = \frac{V}{X_C} = \frac{V}{1/\omega C} = V\omega C$$

The current flowing in the coil $I_L = I_t$ lags, owing to the losses in the coil, behind the voltage by an angle which is somewhat less than 90° . The current I_L has an amplitude

$$I_L = I_t = \frac{V}{z_t} = \frac{V}{\sqrt{r_t^2 + (\omega L)^2}}$$

This current can be represented by the sum of the active (ohmic) component $I_{t.act}$ of the circuit total current I_t , which is in phase with the circuit voltage, and of the reactive component $I_{t.react}$ which lags by 90° behind the circuit voltage. The current I_C flowing through the capacitor and the reactive component $I_{t.react}$ have opposite directions. For instance, if at a given moment the current in the capacitive branch (left) of the circuit is flowing from the top downwards, in the inductive branch (right) it is flowing from the bottom upwards. Therefore, the amplitude of the reactive component $I_{o.react}$ of the total current in the supply circuit, which, in accordance with the Kirchhoff law, presents the sum of currents in the circuit's parallel branches, will be equal to the difference between the amplitudes of the two reactive components:

$$I_{o.react} = I_C - I_{t.react}$$

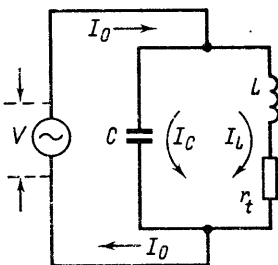


Fig. 2-36. Parallel connection of a circuit to an alternating emf source

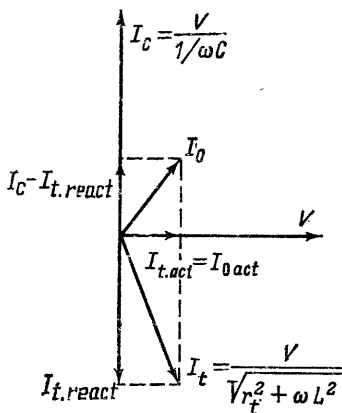


Fig. 2-37. Vector diagram for currents and voltages in a parallel circuit

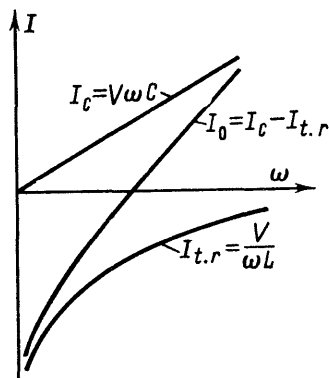
Apart from the reactive component, the active component $I_{0\text{ act}} = I_{t\text{ act}}$ flows through the supply circuit, which compensates for the losses in the oscillatory circuit.

The current flowing through the capacitor increases with frequency (Fig. 2-38):

$$I_C = \frac{V_C}{X_C} = V\omega C$$

If we neglect the resistance of the coil as compared with its reactance which is dozens or even hundreds of times greater than the former, the current $I_{t\text{ react}}$ in the coil varies inversely with frequency:

$$I_{t\text{ react}} = \frac{V}{X_L} = \frac{V}{\omega L}$$



Using curves for the currents I_C and $I_{t\text{ react}}$ it is easy to plot the reactive component of the total current I_0 in the supply circuit. At a frequency where

$$\frac{V}{\omega L} = V\omega C \quad (2-52)$$

Fig. 2-38. Currents in a parallel circuit as functions of frequency

the currents I_C and $I_{t\text{ react}}$ are equal.

As these currents flow in opposite directions, an alternating current circulates in the oscillatory circuit, while the current in the supply line is zero. Consequently, an ideal oscillatory circuit will present, in this case, an infinitely large resistance connected into the supply circuit. The phenomena described above which occur in parallel circuits are called *parallel* or *current resonance*.

In a real parallel circuit with losses, currents in the parallel branches do not compensate each other completely. If the coil resistance r_t is not zero, a small active component $I_{t\text{ act}}$ of the coil current will flow in the supply circuit, and the oscillatory circuit, consequently, will offer only a high resistance to the supply source. From (2-52) it follows that the resonance

frequency

$$\omega = \omega_0 \frac{1}{\sqrt{LC}}$$

At lower frequencies, the current in the inductive branch of the circuit is greater than in the capacitive one (Fig. 2-38) hence the total current $I_0 = I_{t,react}$ in the supply circuit proves to be inductive in nature. At frequencies higher than the resonance frequency, the current of the capacitive branch predominates.

It should be emphasized that the dependence of the total reactance of a parallel circuit on the detuning direction is opposite to that in a series circuit composed of the same elements.

The current in the inductive branch

$$I_t = \frac{V}{\sqrt{r_t^2 + (\omega L)^2}} \quad (2-53)$$

The active component of this current flowing at resonance through the supply circuit is

$$I_{t,act} = I_t \cos \varphi \quad (2-54)$$

where

$$\cos \varphi = \frac{r_t}{\sqrt{r_t^2 + (\omega_0 L)^2}} \quad (2-55)$$

Consequently,

$$I_{or} = I_{t,act} = \frac{V r_t}{r_t^2 + (\omega_0 L)^2} \quad (2-56)$$

from which it follows that, at resonance, the impedance of a parallel circuit is equivalent to the resistance

$$R_{eq} = \frac{V}{I_{or}} = \frac{r_t^2 + (\omega_0 L)^2}{r_t} \quad (2-57)$$

Since r_t is usually dozens or hundreds of times smaller than $\omega_0 L$, and their squares differ dozens of thousands of times, r_t^2 in the numerator can be neglected, thus

$$R_{eq} = \frac{(\omega_0 L)^2}{r_t} \quad (2-58)$$

Taking into account (2-36) and (2-41), this expression can be written as

$$R_{eq} = \frac{\rho^2}{r_t} = Q\rho \quad (2-59)$$

For typical oscillatory circuits used in radio devices, the quality factor Q is of dozens or hundreds, and the effective resistance ρ ranges from dozens to several hundreds of ohms and their equivalent resistance R_{eq} usually ranges from 10^3 to 10^5 ohms.

Also, of some interest is the ratio of the circuit current to the supply current at resonance, i.e. I_t/I_{or} . The current in the oscillatory circuit

$$I_t \approx I_c = V\omega C = \frac{V}{\rho}, \quad (2-60)$$

the supply current

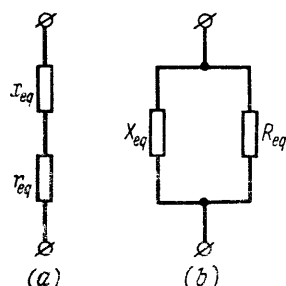
$$I_{or} = \frac{V}{R_{eq}} = \frac{V}{\rho Q} \quad (2-61)$$

and their ratio

$$\frac{I_t}{I_{or}} = Q \quad (2-62)$$

Fig. 2-39. Equivalent circuits of complicated networks

(a) series circuit; (b) parallel circuit



current in a parallel circuit at resonance exceeds that in the supply circuit.

From the above analysis of the phenomena occurring in oscillatory circuits it follows that, depending on its detuning, such a circuit offers to the supply current either a resistance or impedance which can be represented by the combination of an equivalent resistor and a reactive (capacitive or inductive) element connected in series or in parallel.

The method of reducing complicated networks to a simple series (Fig. 2-39a) or parallel (Fig. 2-39b) equivalent circuit is widely used in radio engineering. This is illustrated by the following example which is important for future discussion. We shall find the equivalent parameters of a parallel circuit. Denote the parameters of a series equivalent circuit with the

lower-case letters r_{eq} and x_{eq} , and those of a parallel equivalent circuit, with the upper-case letters R_{eq} and X_{eq} .

Assume that the circuit consists of two parallel branches with the impedances $Z_1 = r_1 + jx_1$ and $Z_2 = r_2 + jx_2$. The admittance of this parallel combination is

$$\frac{1}{Z_{eq}} = \frac{1}{Z_1} + \frac{1}{Z_2}$$

whence,

$$Z_{eq} = \frac{Z_1 Z_2}{Z_1 + Z_2} = \frac{r_1 r_2 - x_1 x_2 + j(x_1 r_2 + x_2 r_1)}{r_1 + r_2 + j(x_1 + x_2)}$$

In practice, the absolute values of x_1 and x_2 are much greater than r_1 and r_2 , respectively, so that all addends in the numerator are much smaller than the product $x_1 x_2$, and, to a high accuracy, we can write that

$$Z_{eq} \approx \frac{-x_1 x_2}{r_t + j(x_1 + x_2)} \quad (2-63)$$

where $r_t = r_1 + r_2$ is the total resistance of the circuit. The resistance r_t in the denominator cannot be neglected because near the resonance the sum $x_1 + x_2$ approaches zero.

Multiplying the numerator and denominator in (2-63) by $r_t - j(x_1 + x_2)$ we obtain

$$Z_{eq} = r_{eq} + jx_{eq} = \frac{-x_1 x_2 r_t}{r_t^2 + (x_1 + x_2)^2} + j \frac{x_1 x_2 (x_1 + x_2)}{r_t^2 + (x_1 + x_2)^2} \quad (2-64)$$

For a parallel circuit

$$-x_1 x_2 = -\omega L \left(-\frac{1}{\omega C} \right) = \frac{L}{C} = \rho^2 \quad (2-65)$$

Dividing the numerator and denominator in (2-64) by r_t^2 and using formula (2-49), we have

$$r_{eq} \approx \frac{1}{1 + Q^2 \left(\frac{2\Delta\omega}{\omega_0} \right)^2} \frac{\rho^2}{r_t} \quad (2-66)$$

$$x_{eq} \approx \frac{\frac{2\Delta\omega}{\omega_0}}{1 + Q^2 \left(\frac{2\Delta\omega}{\omega_0} \right)^2} \frac{\rho^3}{r_t^2} \quad (2-67)$$

The resistive and reactive components of the equivalent impedance of the circuit are plotted in Fig. 2-40a as functions of

the relative detuning $\Delta\omega/\omega_0$. As detuning becomes smaller the resistive component rises sharply, reaching, at resonance, its maximum value, $r_{eq.r} = R_{eq} = \rho^2/r_t = L/Cr_t$. At resonance, the reactive component changes its sign, passing through maxima when the relative detuning $\Delta\omega/\omega_0$ is equal to $\pm 1/2Q$ which, in accordance with (2-15), corresponds to the upper and lower cut-off frequencies of the circuit passband. The absolute value $|x_{eq.\max}|$ is equal to $R_{eq}/2$.

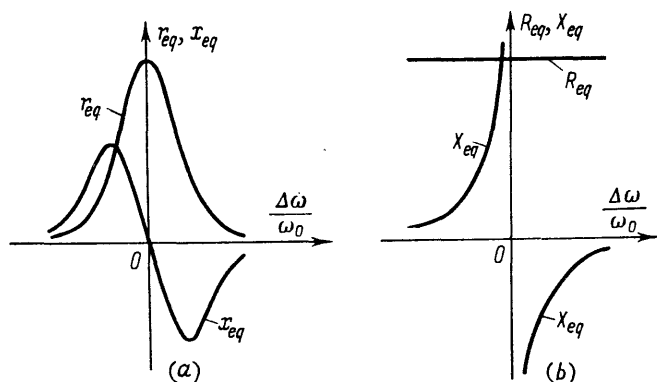


Fig. 2-40. Resistive and reactive components of equivalent impedance of a parallel circuit as functions of relative detuning
(a) for series circuit; (b) for parallel circuit

Let us now determine the relationship between the parameters of a parallel equivalent circuit and the frequency. For small detuning increments $x_1 \approx -x_2 = x_0$; therefore the expression (2-63) can be rewritten as

$$Z_{eq} \approx \frac{x_0^2}{r_t + j(x_1 + x_2)}$$

The admittance of the circuit is equal to the sum of the corresponding parameters of the equivalent circuit; therefore we have

$$\frac{1}{Z_{eq}} \approx \frac{r_t + j(x_1 + x_2)}{x_0^2} = \frac{1}{R_{eq}} + \frac{1}{jX_{eq}} \quad (2-68)$$

By equating the resistive and reactive components in the left and right sides of (2-68) and using formula (2-49) we obtain

$$R_{eq} = \frac{x_0^2}{r_t} = \frac{\omega_0^2 L^2}{r_t} = \frac{1}{(\omega_0 C)^2 r_t} = \frac{\rho^2}{r_t} \quad (2-69)$$

$$\tilde{X}_{eq} = \frac{x_0^2}{x_1 + x_2} \approx \frac{\rho \omega_0}{2\Delta\omega} \quad (2-70)$$

These relationships are plotted in Fig. 2-40b. The curves shown reveal very interesting properties. The resistive component does not depend on frequency and is equal to the cir-

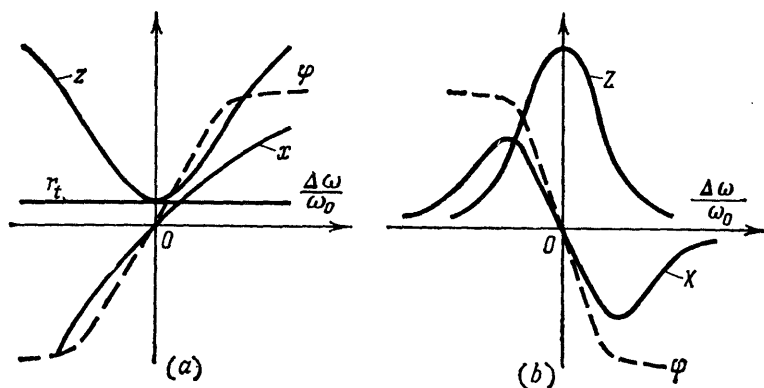


Fig. 2-41. Resistance, reactance, impedance and phase angle between supply current and voltage as functions of relative detuning

(a) for series circuit; (b) for parallel circuit

cuit resistance at resonance. The reactive component has opposite signs on the left and right sides of resonance; it approaches infinity near resonance and then decreases monotonically with increasing detuning. Such relationships between the equivalent parameters and frequency are very convenient for the analysis of processes which take place in radio devices including resonance circuits. For this reason the parallel equivalent circuit will be often used in the further discussion.

In conclusion, let us compare the behaviour of series and parallel circuits as a function of detuning. When detuning is significant a series circuit offers a high reactance to the source current. This reactance is of the capacitive type if the source

frequency is lower than the free oscillation frequency of the circuit, and of the inductive type if it is higher. At resonance, the circuit offers to the source current a resistance which represents the total losses in the circuit elements and is usually of several ohms or fractions of ohms. This behaviour is illustrated in Fig. 2-41a.

A parallel circuit, on the contrary, offers an inductive reactance to the source current whose frequency is lower than the circuit resonance frequency, and a capacitive reactance to the source current whose frequency is higher than the circuit resonance frequency. At resonance the circuit offers to the source current a high resistance (usually measured by several kilohms or dozens of kilohms). This behaviour is illustrated in Fig. 2-41b. The phase shift of the source current with respect to the source voltage as a function of detuning is shown by dash lines in Fig. 2-41.

When the source voltage is constant, the ratio of the current for the detuned circuit to that at resonance is inversely proportional to the ratio of the circuit's equivalent impedances. If we use a parallel equivalent circuit, then

$$\begin{aligned}\frac{I_0}{I_{0r}} = \frac{z_r}{z} &= \frac{\sqrt{\left(\frac{1}{R_{eq}}\right)^2 + \left(\omega C - \frac{1}{\omega L}\right)^2}}{\frac{1}{R_{eq}}} = \\ &= \sqrt{1 + R_{eq}^2 \left(\omega C - \frac{1}{\omega L}\right)^2}\end{aligned}\quad (2-71)$$

By analogy with the expression (2-49), we can write

$$\omega C - \frac{1}{\omega L} \approx \omega_0 C \frac{2\Delta\omega}{\omega_0}$$

Substituting this expression into (2-71) and considering formula (2-41), we obtain

$$\frac{I_0}{I_{0r}} = \sqrt{1 + Q^2 \left(\frac{2\Delta\omega}{\omega_0}\right)^2}\quad (2-72)$$

A parallel circuit is characterized by a passband within which the total current in the source circuit increases less than $\sqrt{2}$ times relative to the current at resonance. The maximum

detuning increment within the passband is found from (2-72):

$$\sqrt{2} = \sqrt{1 + Q^2 \left(\frac{2\Delta\omega_{\max}}{\omega_0} \right)^2}$$

hence, the circuit's bandwidth is

$$2\Delta\omega_{\max} = \frac{\omega_0}{Q} \quad (2-73)$$

From the formulas (2-73) and (2-51) it can be seen that the bandwidths of a parallel and series circuit consisting of the same elements are identical.

It should be pointed out that if a resistor is connected across an oscillatory circuit, some part of the circuit current will flow through this resistor which dissipates power. In accordance with formula (2-58), the connection of the resistor R across the circuit is equivalent to the connection of a resistor with the resistance $r = (\omega L)^2/R$ in series with the circuit coil. Consequently, this resistor inserts additional losses into the circuit, and, thus, lowers the quality factor of the parallel circuit and its resistance at resonance, and widens the circuit passband.

2-9. TRANSIENT PROCESSES IN OSCILLATORY CIRCUITS

In modern radio devices one deals with processes which have a very short duration, lasting several millionths of a second and even less. In spite of their shortness, these processes play a very important part since their duration and character often determine the performance of the radio equipment as a whole. Earlier we became acquainted with transients occurring in circuits with a single energy storing element. Let us examine now transient processes in an oscillatory circuit which includes reactive elements of two types (a coil and capacitor).

First we consider the damping of oscillation which occurs in the circuit when an excitation source is switched off; damping is unavoidable in every real circuit due to the presence of losses in it.

When the source is switched on (the source frequency is assumed to be equal to that of the circuit's free oscillations), sustained oscillations are set up in the circuit, during which the

energy ΔW_{react} stored in the circuit coil changes in a half-cycle from $I^2L/2$ to zero. In accordance with expression (2-23), the power dissipated during this transition across the circuit resistance is

$$\Delta W_{act} = \frac{1}{2} I^2 r \frac{T}{2}$$

The ratio of the energy lost to the reactive energy stored in the circuit during oscillations will be

$$\frac{\Delta W_{act}}{\Delta W_{react}} = \frac{r}{2L} T$$

This quantity divided by the oscillation period is called *circuit damping factor*:

$$\alpha = \frac{r}{2L} \quad (2-74)$$

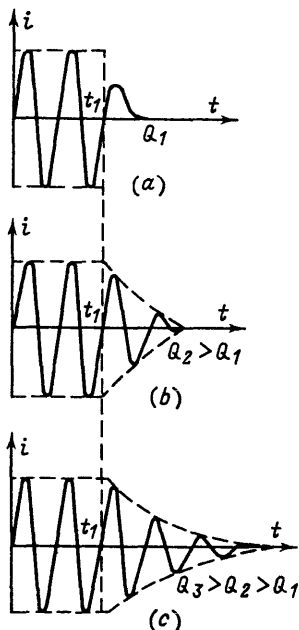
It is clear that this factor characterizes the time rate at which the oscillation amplitude is decreasing when the excitation source is switched off

If the losses in the circuit are so great that all the energy stored in it is lost during a half-cycle, only one current pulse occurs in the circuit after the source has been switched off (Fig. 2-42a). If, however, losses in the circuit are smaller, oscillations with decreasing amplitude are set up in the circuit, the number of cycles appearing in the circuit until the oscillation has ceased

Fig. 2-42. Current in oscillatory circuits having different Q -factors; source is switched off

will be the greater, the smaller the circuit damping factor (Fig. 2-42b and c). The duration of the damping process can be evaluated by a quantity reciprocal of the damping factor and called time constant

$$\tau = \frac{2L}{r} \quad (2-75)$$



In time $t=\tau$ after the source has been switched off, the oscillation amplitude decreases by 2.7 times, in $t=2\tau$, by $2.7^2=7.3$ times, and in $t=3\tau$, by $2.7^3\approx 20$ times.

In the last case the damping process can be considered practically completed, i.e. the oscillation has ceased.

As $Q=\omega_0 L/r$, so $L/r=Q/\omega_0$. Substituting this ratio into the expression for the time constant (2-75), we obtain

$$\tau = \frac{2Q}{\omega_0} = \frac{Q}{\pi f_0} \quad (2-76)$$

In other words, the transient process duration t_{st} will be the longer, the higher the circuit quality factor. For practical purposes it can be assumed that

$$t_{st} = 3\tau = \frac{3Q}{\pi f_0} \approx \frac{Q}{f_0} \quad (2-77)$$

Insofar as the quality factor is related to the circuit bandwidth by the following equation

$$Q = \frac{\omega_0}{2\Delta\omega_{\max}} = \frac{f_0}{2\Delta f_{\max}}$$

it follows that

$$t_{st} = \frac{3}{2\pi\Delta f_{\max}} \approx \frac{1}{2\Delta f_{\max}} \quad (2-78)$$

hence, the larger the bandwidth of a circuit, the sooner the oscillation in it decays.

When the alternating voltage source connected to an oscillatory circuit is switched on, oscillation in the latter will rise slowly (Fig. 2-43). The rate of its rise is limited by the losses in the total resistance of the circuit. Evidently, the less the energy stored in the circuit under its steady-state condition and the smaller the amount of the source power lost in the total resistance of the circuit, the quicker the circuit reactive elements accumulate the energy. In other words, the time required for oscillation in a circuit to reach its steady-state amplitude and that to decay are both determined by the same parameters of the circuit, i.e. by its time constant.

All the above helps the understanding of the processes which take place in an oscillatory circuit when a pulse of high frequency voltage is applied to it (Fig. 2-44a). Since the amplitude of oscillations in the circuit rises and decreases gradually, the shape of the pulse is distorted. These distortions will be small if the pulse duration is much greater than that of the

transients in the circuit (Fig. 2-44b). If these durations, however, are of the same order, distortion becomes considerable (Fig. 2-44c). Finally, if the duration of the transients is greater than that of the pulse, the amplitude of oscillation starts to decrease even not reaching the amplitude of the pulse by the end of the latter (Fig. 2-44d). The duration of the distorted pulse, in this case, will be much longer than that of the input pulse, and its amplitude, as mentioned above, will not reach the maximum possible value.

It follows that in order to minimize the pulse distortion, it is necessary to employ a circuit with the least possible t_{st} , i.e. a circuit with a sufficiently wide passband and a low quality factor. It should be

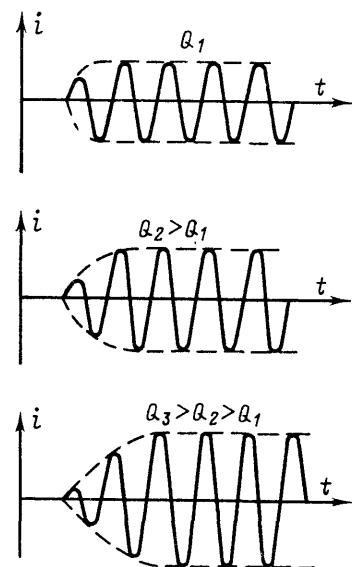


Fig. 2-43. Current in oscillatory circuits having different Q -factors; source is switched on

borne in mind, however, that lower circuit Q -factors result in lower effective resistance of the circuit at resonance, which, of course, is undesirable. Therefore, the value of the circuit Q -factor is usually selected so that the oscillation amplitude can reach its maximum value for the time τ_{in} . Whence, on the basis of the expressions (2-76) through (2-78), the circuit Q -factor will be $Q \leq 3\tau f_0$, and the circuit bandwidth $2\Delta f_{max} \geq 1/3\tau$.

These conditions often make it necessary to employ circuits with a very wide passband. For instance, in radars pulses

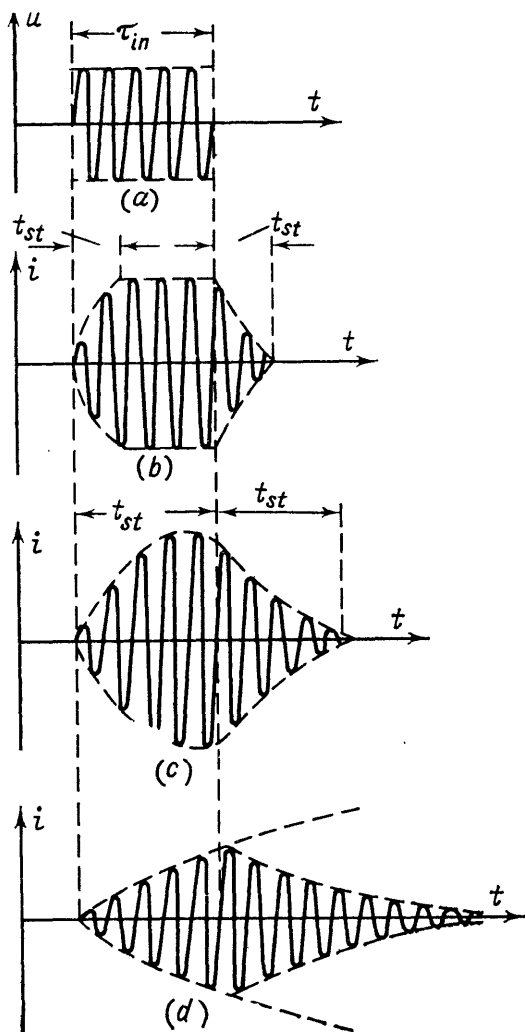


Fig. 2-44. Short high-frequency voltage pulse acting on oscillatory circuits with different Q -factors

having duration of about a microsecond ($\tau = 1 \times 10^{-6}$ s) are used. In order to reduce distortions of such a pulse to a minimum the circuits of the radar receiver must have a bandwidth not less than $2\Delta f_{\max} = 1/3 \times 10^{-6} \approx 0.3$ MHz.

2-10. PARALLEL CIRCUITS

WITH REACTIVE ELEMENTS OF BOTH TYPES IN ONE OF THEIR BRANCHES

So far we have examined parallel circuits which include a capacitor in one branch, and a coil in the other. Such circuits we shall call *circuits of the first type*. In radio engineering, however, parallel circuits whose branches include a coil and capacitor connected in series are often used. A circuit including

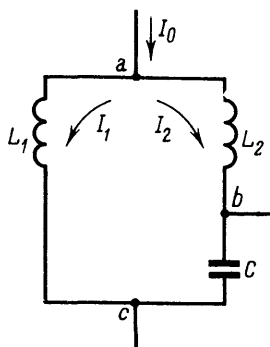


Fig. 2-45. Circuit of second type

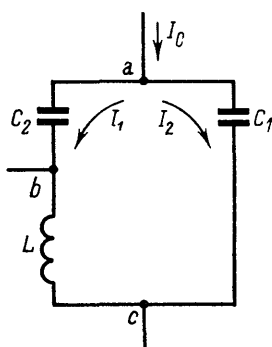


Fig. 2-46. Circuit of third type

a coil in each of its two branches (Fig. 2-45) we shall call *circuit of the second type* (it can be converted into a circuit of the first type by connecting the supply source to the points *b* and *c*). A circuit with a capacitor in each of its two branches (Fig. 2-46) is called *circuit of the third type* (it can be converted into one of the first type by connecting the supply source to the points *b* and *c*).

An interesting feature of these circuits is that they exhibit resonance at two different frequencies. In the case when

$$\omega_1 L_2 = \frac{1}{\omega_1 C} \quad (2-79)$$

series resonance occurs in the right branch of the second-type circuit (Fig. 2-45). For this reason the right branch presents a small resistance and almost all the current flows through it. Thus, this branch practically determines the circuit equivalent impedance which has the resistive character.

At a lower frequency, at which the sum of all circuit reactances is zero, i.e. when

$$\omega_2 L_1 + \omega_2 L_2 - \frac{1}{\omega_2 C} = 0 \quad (2-80)$$

parallel resonance takes place in the circuit and it presents a high resistance. Since the resistances of the circuit branches are in this case smaller compared to their reactances, the currents in the right and left branches (I_2 and I_1) are phase shifted relative to the voltage across the circuit by $+90^\circ$ and -90° respectively. Therefore, a small active current flows in the circuit.

At frequencies close to the parallel resonance frequency, a change in the circuit equivalent impedance corresponds to that in the reactance of a parallel circuit, whereas at frequencies close to the series resonance frequency it corresponds to a change in the reactance of a series circuit. This can be explained with the aid of Fig. 2-47 where the reactances of the circuit are plotted against frequency. In Fig. 2-47a the equivalent reactance $X_2 = X_{L2} - X_C$ of the circuit's right branch (see Fig. 2-45) is plotted, while Fig. 2-47b shows its admittance $y_2 = -1/X_2$. In Fig. 2-47c this admittance is added graphically to the admittance $y_1 = -1/X_{L1}$ of the left branch of the circuit. In Fig. 2-47d the circuit equivalent reactance $X_{eq} = -1/(y_2 + y_1)$ is plotted.

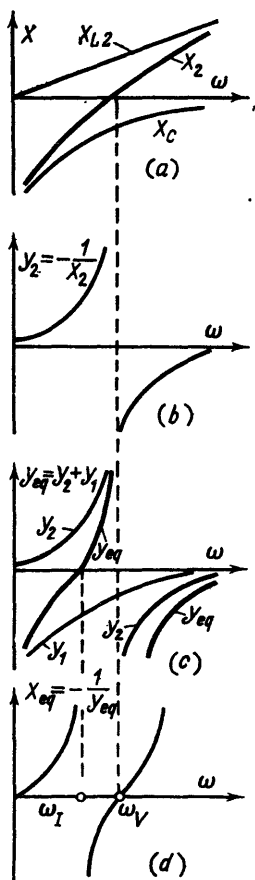


Fig. 2-47. Reactances and admittances of the circuit of second type as functions of frequency

Let us find the relation between the two resonance frequencies. From expression (2-79)

$$\omega_1 = \frac{1}{\sqrt{L_2 C}} \quad (2-81)$$

and from (2-80) we have

$$\omega_2 = \frac{1}{\sqrt{(L_1 + L_2) C}} \quad (2-82)$$

From the last two equations it follows that

$$\omega_1 = \omega_2 \sqrt{1 + \frac{L_1}{L_2}} \quad (2-83)$$

A circuit of the second type (Fig. 2-45) can be obtained from that of the first type by connecting one of the two supply lines to a tap from its coil. The ratio of the inductance which forms the inductive branch of such a circuit to that of the whole coil we shall call *circuit tapping factor*:

$$p = \frac{L_1}{L_1 + L_2} \quad (2-84)$$

Formula (2-83) can be now rewritten as

$$\omega_1 = \omega_2 \sqrt{\frac{1}{1-p}} \quad (2-85)$$

At parallel resonance, when the currents in the two circuit branches are approximately equal, the tapping factor characterizes the ratio of the voltage across the circuit terminals, i.e. the potential difference between the points *a* and *c*, to that across the whole coil, i.e. the potential difference between the points *b* and *c* (Fig. 2-45):

$$\frac{V_{a-c}}{V_{b-c}} = \frac{I_t L_1}{I_t (L_1 + L_2)} = p \quad (2-86)$$

The power consumed by the circuit at resonance when current I_t flows through it, is

$$P = \frac{1}{2} I_t^2 r_t \quad (2-87)$$

where r_t is the total circuit resistance. The voltage produced across the terminals *a* and *c* by the current I_t (without taking

into account the circuit total resistance) is

$$V_t = I_t \omega_2 L_1$$

The power dissipated in the circuit can also be calculated from the following formula

$$P = \frac{1}{2} \frac{V_t^2}{R_{eq2}} = \frac{I_t^2 (\omega_2 L_1)^2}{2 R_{eq2}} \quad (2-88)$$

Comparing expressions (2-87) and (2-88), we obtain

$$R_{eq2} = \frac{\omega_2^2 L_1^2}{r_t} = p^2 \frac{(\omega_2 L)^2}{r_t} \quad (2-89)$$

where $L = L_1 + L_2$.

The factor $(\omega_2 L)^2 / r_t$ represents the equivalent resistance R_{eq} of the circuit when the latter is used as a circuit of the first type. Consequently

$$R_{eq2} = p^2 R_{eq} \quad (2-90)$$

This result is of great practical importance: if one of the circuit terminals is made in the form of a feeler with a clip at its end, the circuit equivalent resistance can be reduced to any desired value by moving this terminal along the coil. Such method of varying the circuit equivalent resistance is widely used in radio transmitters.

Let us now examine the oscillatory circuits of the third type (see Fig. 2-46). These also have two resonances: series resonance in the left branch, when

$$\omega_1 L = \frac{1}{\omega_1 C_2} \quad (2-91)$$

and parallel resonance, when the reactances of both branches are equal, i.e. when

$$\frac{1}{\omega_2 C_1} = \omega_2 L - \frac{1}{\omega_2 C_2} \quad (2-92)$$

Let us find the relation between the two resonance frequencies. From expression (2-91) we have

$$\omega_1 = \frac{1}{\sqrt{LC_2}} \quad (2-93)$$

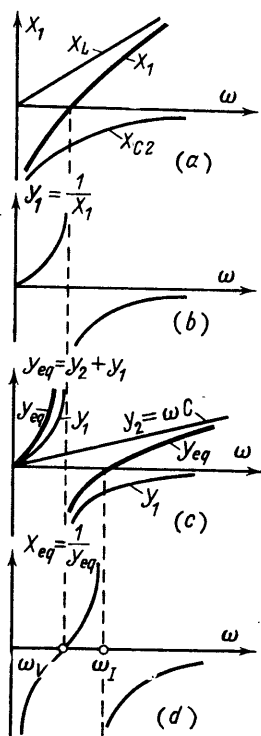


Fig. 2-48. Reactances and admittances of the circuit of third type as functions of frequency

In Fig. 2-48a the sum of reactances $X_1 = -1/\omega C_2 + \omega L$ of the circuit's left branch is plotted; Fig. 2-48b gives the susceptance $y_1 = -1/X_1$ of the circuit's left branch, Fig. 2-48c represents the sum of the susceptances of both branches $y_{eq} = y_2 + y_1$; and, finally, Fig. 2-48d shows the circuit equivalent reactance $X_{eq} = -1/y_{eq}$.

These plots, as well as formula (2-97), indicate that series resonance in the third-type circuits, as distinct from the second-type circuits, occurs at a lower frequency than parallel resonance. Near series resonance the circuit reactance varies in the same way as in the case of a series circuit, and near pa-

and from expression (2-92)

$$\omega_2 = \frac{1}{\sqrt{L \frac{C_1 C_2}{C_1 + C_2}}} \quad (2-94)$$

The last two equations give

$$\omega_2 = \omega_1 \sqrt{\frac{C_1 + C_2}{C_1}} \quad (2-95)$$

The tapping factor for a circuit of the third type defined as the ratio of the resonance voltage across the terminal points *a* and *c* (Fig. 2-46) to the voltage across the two capacitors, is

$$p = \frac{I_t \frac{1}{\omega C_1}}{I_t \left(\frac{1}{\omega C_1} + \frac{1}{\omega C_2} \right)} = \frac{C_2}{C_1 + C_2} \quad (2-96)$$

Thus,

$$\omega_2 = \omega_1 \sqrt{\frac{1}{1-p}} \quad (2-97)$$

The behaviour of the circuit equivalent reactance as a function of circular frequency can be determined from that of the equivalent reactances of circuit elements (Fig. 2-48). In Fig. 2-48a the sum of reactances $X_1 = -1/\omega C_2 + \omega L$ of the

parallel resonance in the same way as in the case of a parallel circuit.

As the meaning of the tapping factor for a circuit of the third type is the same as that for the second type, the circuit equivalent resistance at resonance is $R_{eq3} = p^2 R_{eq}$. In the third-type circuit, however, it is less convenient to control the circuit equivalent resistance by changing the tapping factor than in the second-type circuit, since in the first case both capacitances C_2 and C_1 must be changed which also leads to a change in the circuit resonance frequency. For this reason, it is necessary to control both capacitances C_1 and C_2 simultaneously while keeping the total circuit capacitance constant (if the capacitance C_1 is decreased, C_2 must be increased). In spite of this disadvantage, the circuits of the third type are often employed in modern radio equipment.

3-1. OSCILLATIONS IN TWO COUPLED CIRCUITS

In radio equipment an oscillatory circuit is usually coupled with other networks which receive some part of energy supplied to the oscillatory circuit by an external source. A network

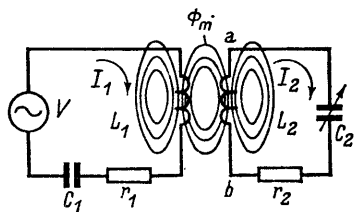


Fig. 3-1. Two inductively coupled circuits

coupled with an oscillatory circuit is often an oscillatory circuit itself. We shall call *primary* the circuit whose oscillations are set up by the external source, and *secondary* the circuit to which some part of energy from the primary circuit is transmitted.

Two circuits can be coupled in different ways. For instance,

if the coils of the two circuits are placed close to each other (Fig. 3-1) a part of the alternating magnetic flux produced by the current in the coil of the primary circuit will thread through the turns of the coil in the secondary circuit inducing an alternating emf in it. The value of this emf can be determined by the law of electromagnetic induction:

$$E_{12} = -w_2 \frac{\Delta \Phi_m}{\Delta t} \quad (3-1)$$

where Φ_m is the common or mutual magnetic flux (instantaneous value) which threads both coils, and w_2 denotes the number of turns in the secondary-circuit coil.

The mutual magnetic flux which links the two coils is proportional to the current through the primary-circuit coil. If the current changes by the sine wave law, then $w_2 \Phi_m =$

$=MI_1\sin\omega t$, where M is the proportionality factor called *mutual inductance*.

The emf induced in the secondary-circuit coil is proportional to the rate of flux change; thus, in accordance with expression (2-4), $E_{12} = -M\omega I_1\sin(\omega t + 90^\circ)$. The amplitude of the induced emf is

$$E_{12} = \omega MI_1 \quad (3-2)$$

The quantity ωM is measured in ohms, and is called *mutual reactance (impedance)*:

$$\omega M = X_m \quad (3-3)$$

As the mutual inductance and, consequently, the mutual reactance of the coupled circuits rise, the emf induced in the secondary circuit will also rise.

Under the action of the induced emf a current I_2 begins to flow in the secondary circuit. This current, passing through the coil L_2 , produces an alternating magnetic flux part of which threads through the turns of the primary circuit coil and, in its turn, induces in it an emf. If the current in the secondary circuit is shifted by an angle φ relative to the induced emf, i.e. $i_2 = I_2 \sin(\omega t + 90^\circ + \varphi)$ then the emf induced in the primary circuit is

$$E_{21} = -\omega MI_2 \sin(\omega t + 180^\circ + \varphi)$$

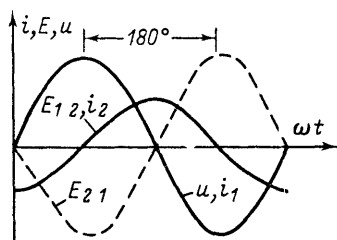


Fig. 3-2. Currents and voltages in coupled circuits tuned to the same frequency

In Fig. 3-2 currents, voltages and emf's in the coupled circuits are plotted as functions of ωt for the case when both circuits are tuned in resonance at the angular frequency ω and the phase shift φ is equal to zero. From the plots given it can be seen that, under the influence of the secondary circuit an emf E_{21} is induced in the primary circuit, which is directed opposite to the source voltage u producing current in the primary circuit. Therefore, the current in the primary circuit decreases.

The effect is such as if the resistance of the primary circuit were increased due to coupling. This is explained by the fact that energy fed from the external source to the primary circuit is consumed not only in it, but is also partially transmitted into the secondary circuit. The consumption by the secondary circuit of power transmitted from the primary one is equivalent to connecting a resistor into the primary circuit, which dissipates the same amount of power as that consumed by the secondary circuit.

Therefore, it is said that the secondary circuit inserts ohmic losses in the primary one. This reduces the Q -factor of the primary circuit as well as its effective resistance (at resonance).

Let us determine the value of the *insertion loss*. The current in the secondary circuit

$$I_2 = \frac{E_{12}}{Z_2} = \frac{I_1 \omega M}{r_2} \quad (3-4)$$

where Z_2 is the impedance of the secondary circuit, which is equal at resonance to its total resistance r_2 . The power extracted by the secondary circuit from the primary one is

$$P_2 = \frac{1}{2} I_2^2 r_2 = \frac{1}{2} I_1^2 \frac{\omega^2 M^2}{r_2} \quad (3-5)$$

This power is consumed as if it were dissipated in a resistor Δr inserted into the primary circuit, thus we can write

$$P_2 = \frac{1}{2} I_1^2 \Delta r = \frac{1}{2} I_1^2 \frac{\omega^2 M^2}{r_2} \quad (3-6)$$

thus, the insertion loss is characterized by the resistance

$$\Delta r = \frac{\omega^2 M^2}{r_2} = \frac{X_m^2}{r_2} \quad (3-7)$$

The total power supplied by the external source in the primary circuit

$$P_1 = \frac{1}{2} I_1^2 (r_1 + \Delta r) \quad (3-8)$$

Considering the power transmitted to the secondary circuit to be useful, we introduce the notion of the *primary-circuit efficiency* η which shows what part of the total power supplied to it is transmitted to the secondary circuit. From formulas

(3-6) and (3-8), we have

$$\eta = \frac{P_2}{P_1} = \frac{\Delta r}{r_1 + \Delta r} = \frac{1}{1 + \frac{r_1}{\Delta r}} \quad (3-9)$$

This expression shows that the higher the primary-circuit efficiency, the greater the ratio of the insertion loss to the total resistance of the primary circuit itself. This, however, does not mean that the absolute value of the power transmitted to the secondary circuit will rise continuously with increasing the extent of coupling between the two circuits and insertion losses. For a fixed source voltage, the current in the primary circuit decreases as the insertion loss is increased (Fig. 3-3):

$$I_1 = \frac{V}{r_1 + \Delta r} = \frac{V}{r_1} \frac{1}{1 + \frac{\Delta r}{r_1}} \quad (3-10)$$

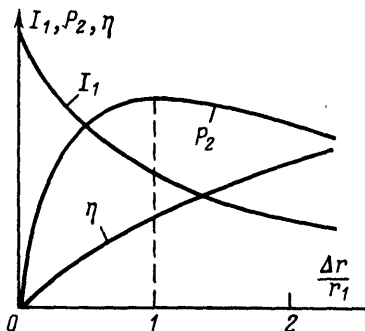


Fig. 3-3. Power relationships for two coupled circuits as functions of resistance inserted into the primary circuit

The power dissipated in the inserted resistor Δr (transmitted to the secondary circuit) is

$$P_2 = \frac{1}{2} I_1^2 \Delta r = \frac{1}{2} \left(\frac{V}{r_1} \right)^2 \frac{1}{\left(1 + \frac{\Delta r}{r_1} \right)^2} \Delta r \quad (3-11)$$

The power P_2 which is a product of two quantities one of which is continuously decreasing and the other increasing, reaches a maximum when $\Delta r/r_1 = 1$ (Fig. 3-3). When this occurs the power transmitted to the secondary circuit

$$P_2 = \frac{P_1}{2} = \frac{1}{8} \frac{V^2}{r_1} \quad (3-12)$$

and the primary circuit efficiency η_{opt} is equal to 1/2.

Under certain conditions in two coupled circuits each having the same frequency of free oscillation a resonance can appear not only at this frequency, but also at two other frequencies.

In fact, at some frequency which is somewhat lower than that of free oscillation the impedances of the two circuits are of the capacitive type. The emf E_{12} induced in the secondary circuit lags behind the current I_1 in the primary circuit by 90° (Fig. 3-4); since the impedance of the secondary circuit is of the capacitive type, the current I_2 in this circuit leads the emf E_{12} by angle φ . The emf E_{21} induced in the

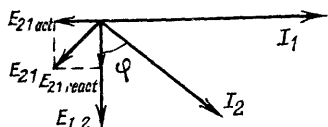


Fig. 3-4. Vector diagram for currents and voltages in two coupled circuits; source frequency is lower than free-oscillation frequency

primary circuit due to the flux produced by the secondary circuit lags behind the current I_2 in the secondary circuit by 90° . This emf can be resolved into two components, one of which E_{21act} is directed opposite to the vector of the current in the primary circuit, and the other $E_{21react}$ lags behind it by 90° .

Consequently, the first component tends to decrease the current in the primary circuit, which is equivalent to an increase of the total resistance of this circuit. The second component tends to increase the emf induced in the coil L_1 , which lags behind the current by 90° (see Section 2.2); the latter effect is equivalent to an increase of the inductance of the coil. Therefore, it is said that when the impedance of the secondary circuit is of the capacitive type, additional resistance and inductive reactance are inserted into the primary circuit. Evidently, these components will be the greater, the closer the coupling between the two circuits. If the coupling is tight enough and the coupled circuits are continuously detuned relative to the source frequency, then at a certain detuning the inserted inductive reactance will neutralize the capacitive reactance of the primary circuit. Under these conditions the primary circuit offers a small pure resistance to the source current so that a series resonance occurs in it. Owing to this, the currents in the primary and secondary circuits increase steeply.

It is easy to prove that a similar situation will take place also at a frequency somewhat higher than that of the circuit free oscillation. In this case, the impedances of the two circuits are of the inductive type. Therefore, the current I_2 in the secondary circuit lags behind the induced emf E_{12} by angle φ

(Fig. 3-5). The emf $E_{2,1}$ induced by this current in the primary circuit again has two components. The first component $E_{2,1 \text{ act}}$ tends to reduce the current in the primary circuit; this effect is equivalent to an increase of the total resistance of the primary circuit. The second component $E_{2,1 \text{ react}}$ leads the current by 90° , thus reduces the voltage across the capacitor of the primary circuit. The same effect would be produced by connecting an additional capacitor in series with that of the circuit. Therefore, it is said that the secondary circuit inserts a resistance and a capacitive reactance into the primary circuit. If the coupling is tight enough and the coupled circuits, as before, are continuously detuned relative to the source frequency, then at a certain detuning the inserted capacitive reactance will neutralize the inductive reactance of the primary circuit. Under these conditions the primary circuit again offers a small pure resistance to the source current so that another series resonance occurs in it and the currents in both circuits rise sharply.

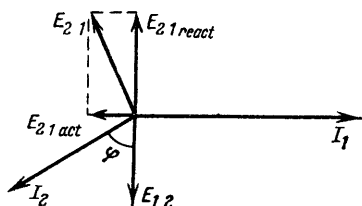


Fig. 3-5. Vector diagram for currents and voltages in two coupled circuits; source frequency is higher than free-oscillation frequency

Thus, the system of two identical circuits coupled sufficiently tightly proves to have three resonance frequencies: one is the frequency of their free oscillation, the other two are somewhat lower and higher than the first one, respectively. The last two resonance frequencies are also called *coupling frequencies*. The tighter the coupling between the two circuits, the greater the reactance inserted by the secondary circuit into the primary. Therefore, the reactances neutralize each other at a greater detuning of the primary circuit and, consequently, the coupling frequencies differ more from the frequency of the circuit free oscillation.

If the coupling is below a certain critical value, the inserted reactances are not sufficient to neutralize the circuit inherent reactances, and resonance will appear only at the circuit free-oscillation frequency. Therefore, when the coupling is less than critical, the resonance characteristic (curve 1 in

Fig. 3-6) resembles that of a single oscillatory circuit. When the coupling is higher than critical, the resonance characteristic (curve 2) assumes the form of a double-humped curve whose maximums correspond to the coupling frequencies ω_1 and ω_2 . If the coupling is still tighter, the coupling frequen-

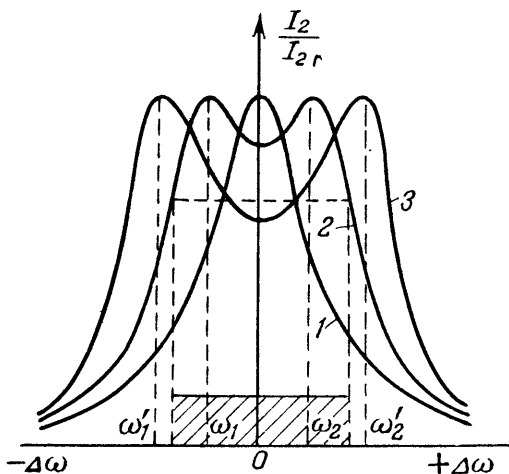


Fig. 3-6. Resonance characteristics of two coupled circuits for different values of coupling coefficient

cies ω'_1 and ω'_2 are located even farther from each other (curve 3).

Let us now determine the coupling frequencies. With the current I_1 in the primary circuit, the emf E_{12} induced in the secondary circuit is ωMI_1 . Then the current in the secondary circuit will be

$$I_2 = \frac{E_{12}}{Z_2} = \frac{\omega MI_1}{Z_2}$$

The emf induced in the primary circuit by the current I_2 can be found from the formula

$$E_{21} = \omega MI_2 = \frac{\omega^2 M^2 I_1}{Z_2}$$

The impedance inserted into the secondary circuit by the pri-

mary one is

$$\frac{E_{21}}{I_1} = \frac{\omega^2 M^2}{Z_2} = \frac{\omega^2 M^2}{r_2 + jx_2}$$

By multiplying the numerator and denominator of the last equation by $r_2 - jx_2$, we obtain

$$\frac{E_{21}}{I_1} = \Delta r + j\Delta x = \frac{\omega^2 M^2 r_2}{r_2^2 + x_2^2} - j \frac{\omega^2 M^2 x_2}{r_2^2 + x_2^2}$$

from which formulas for the active (ohmic) and reactive components of the inserted impedance can be determined:

$$\begin{aligned} \Delta r &= \frac{\omega^2 M^2}{r_2^2 + x_2^2} r_2 = \frac{X_m^2}{z_2^2} r_2 \\ \Delta x &= - \frac{\omega^2 M^2}{r_2^2 + x_2^2} x_2 = - \frac{X_m^2}{z_2^2} x_2 \end{aligned}$$

At the coupling frequencies

$$x_1 = -\Delta x = \frac{X_m^2}{r_2^2 + x_2^2} x_2$$

The total resistance, i.e. loss in coupled circuits employed in radio equipment, is usually small, much smaller than the reactances of their reactive elements; therefore, with accuracy sufficient for practical purposes we can assume that $z_2^2 = r_2^2 + x_2^2 \approx x_2^2$. If both circuits are identical, $x_1 = x_2 = x$. Then the condition for resonance assumes the following form: $x = X_m^2/x$, whence $x = \omega L - 1/\omega C = \pm X_m$, or

$$\omega L \left(1 - \frac{1}{\omega^2 LC} \right) = \pm X_m$$

Since $1/LC = \omega_0^2$, we get $1 - (\omega_0^2/\omega^2) = \pm M/L$. Now we can write the formula for the coupling frequencies:

$$\omega_{1,2} = \frac{\omega_0}{\sqrt{1 \pm \frac{M}{L}}} \quad (3-13)$$

The ratio M/L is called *coefficient of coupling*. If we multiply the numerator and denominator of this expression by ωI_1 , it will turn out that the coupling coefficient expresses

the ratio of the emf induced in the secondary circuit to the voltage across the induction coil of the primary circuit. Designating this ratio by k , i.e.

$$\frac{M}{L} = k \quad (3-14)$$

from formula (3-13) we obtain the expression for the first coupling frequency which we shall call *lower resonance frequency*,

$$\omega_1 = \omega_l = \frac{\omega_0}{\sqrt{1+k}} \quad (3-15)$$

and for the second, *upper resonance frequency*,

$$\omega_2 = \omega_u = \frac{\omega_0}{\sqrt{1-k}} \quad (3-16)$$

These formulas indicate that the greater the coupling coefficient, the greater is the difference between the two coupling frequencies and the free frequency of the circuits, the lower coupling frequency being always lower, and the upper, always higher than the circuit's free frequency (Fig. 3-7).

It should be borne in mind that formulas (3-13) through (3-16) are approximate, as they are derived on the assumption that the coupled circuits are lossless. For large coupling coefficients and high Q -factors of the circuits these formulas provide reasonable accuracy.

If the source frequency is lower or higher than both coupling frequencies, the current in the secondary circuit decreases more rapidly than in the case of a single resonance circuit. In other words, the selectivity of the system of two coupled circuits, i.e., its ability to separate signals of desired frequencies and suppress those of all other frequencies, is higher than that of a single resonance circuit. This fact is widely used in practice. Another advantage of coupled circuits is their ability to pass, without appreciable distortions, the signals within a certain frequency band (the shaded area for the curve 2 in Fig. 3-6).

In practice one has often to deal with a system of two coupled circuits which differ in their free frequency. Thus, it is important to know the resonant properties of such circuits. First, let us determine a band within which their coupling

frequencies lie. Following the way of reasoning used in the case of two identical coupled circuits, we can plot the vector diagrams for the system considered.

It is easy to understand that in case the source frequency is lower than the free frequency of either of the circuits, the vector diagram for the system will have the form similar to that for two identically tuned circuits (see Fig. 3-4). Thus, under this condition the system exhibits resonance provided the coupling is sufficiently tight. The tighter the coupling between the

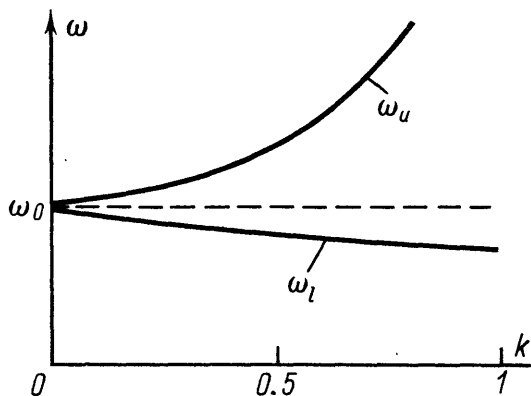


Fig. 3-7. Free-oscillation frequencies of two coupled circuits as functions of coupling coefficient

two circuits, the greater the neutralizing reactance inserted into the primary circuit by the secondary one. Correspondingly, the greater will be the difference between the resonance frequency of the coupled circuits and their free frequencies.

Suppose that the source frequency lies between the free frequencies of the two circuits, and that the primary circuit is tuned to the lower frequency, then the vector diagram will have the same form as in Fig. 3-4. However, in this case the source sees the primary circuit as a capacitive reactance which is further increased due to the reactance of the same type inserted by the secondary circuit, and thus resonance cannot occur.

Similarly, it can be shown that a second resonance is possible at a frequency higher than the free frequency of either of the two circuits. On the other hand, if the free frequency of the

primary circuit is higher than that of the secondary circuit and the source frequency lies between the two, no resonance will occur.

Thus, the system of two non-identical coupled circuits can have two resonance frequencies (coupling frequencies): one is lower than the lower of the two free frequencies, and the

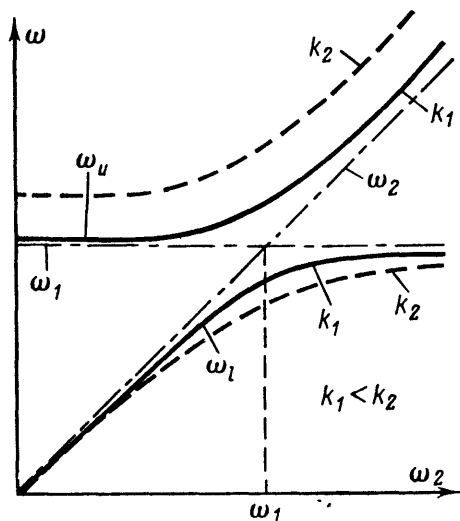


Fig. 3-8. Free-oscillation frequencies of coupled circuits as functions of tuning for different coupling coefficients

other is higher than the higher of them. The closer the coupling, the greater is the difference between the two resonance frequencies.

Figure 3-8 shows the results obtained in calculating the resonance frequencies for two coupled circuits with the coupling coefficient k_1 and different free frequencies. The curves given indicate that as the secondary-circuit frequency rises while the primary-circuit frequency is fixed, the two resonance frequencies also grow, one of them remaining below the frequency of the primary circuit, and the other, above that of the secondary. If we calculate or measure the same dependencies for a greater coupling coefficient k_2 , we shall obtain identical

curves (dashed) spaced farther from each other, i.e., the resonance frequencies will be also spaced farther from each other.

Previously we examined the general properties of coupled circuits for the particular case of the circuits with mutual inductance coupling. These basic properties are also applied to any type of coupled circuits with the only difference that the expressions for the mutual impedance, coupling coefficient and coupling (resonance) frequencies will be somewhat modified.

In Fig. 3-9a circuits with direct inductive coupling are shown. In this case, a part of the voltage across the coil L_1 , which is produced across L_m , is fed to the secondary circuit, thus exciting electrical oscillations in it. In Fig. 3-9b the two circuits are coupled with the help of the capacitor C_m connected into their common branch (internal capacitive coupling). Two parallel branches are formed for the currents of the primary circuit: the coupling capacitor C_m in one branch and the elements L_2 , C_2 , r_2 in the other. The greater the reactance of the coupling capacitor, i.e., the smaller its capacitance, the greater amount of the current in the primary circuit will be branched off to the secondary and, therefore, the tighter will be the coupling. Another type of capacitive coupling (external

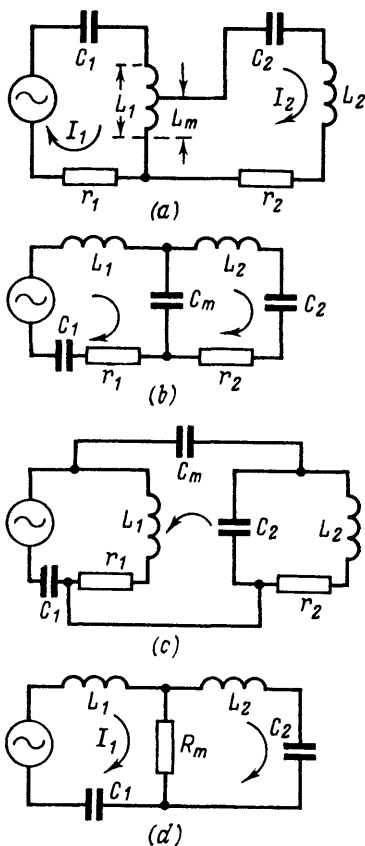


Fig. 3-9. Types of coupling between circuits

capacitive coupling) between circuits is shown in Fig. 3-9c. In this case the current in the secondary circuit will be the greater, the smaller the reactance of the coupling capacitor, i.e., the greater its capacitance. Fig. 3-9d shows a system with conductive coupling between the circuits, where the resistor R_m is common for both circuits. In almost all of these systems the element which is common for both circuits serves as a coupling element. The voltage across the secondary circuit is equal to the product of the coupling element reactance (or resistance) and the current flowing through it. The only exception is the system given in Fig. 3-9c, where the coupling element is a capacitor connected between the two circuits.

3-2. TUNING A SYSTEM OF TWO COUPLED CIRCUITS

The purpose of tuning a system of coupled circuits is usually to transmit as much power as possible to the secondary circuit, i.e., to obtain in it the maximum current or a required passband while providing the maximum possible selectivity. The current in the secondary circuit can be increased by several methods. For instance, by tuning the primary circuit to resonance at one of the coupling frequencies, which will result in an increase of current in the primary as well as in the secondary circuit. Such a case is called *first particular resonance*. The same effect can be produced by tuning the secondary circuit to resonance at one of the coupling frequencies. The latter case is called *second particular resonance*.

In both cases the type and extent of coupling between the circuits can be chosen arbitrarily; therefore, it may happen that a very small amount of power will be transmitted to the secondary circuit. To transmit the maximum possible power to the secondary circuit, it is necessary, besides tuning, to choose the optimal coupling. Thus, two conditions must be met for the maximum power to be transmitted from the primary circuit to the secondary: (1) the reactance inserted into the primary circuit must be equal and opposite in sign to its own reactance, and (2) the resistance inserted into the primary circuit must be equal to its own total resistance.

It is said that a *common resonance* takes place in the system which has been adjusted as described above. In practice this

is accomplished through a number of successive adjustments. For instance, a loose coupling is first established between the circuits and the primary circuit is tuned so that the maximum current is achieved in the secondary. Then the extent of coupling is increased, the tuning is repeated, and the new maximum value of current in the secondary circuit is noted. This operation is repeated until a further increase of coupling results in a decrease of the current in the secondary circuit as compared to the preceding maximum values. The same results can be obtained by tuning the secondary circuit and selecting an optimum coupling.

It is too complicated and inconvenient to obtain the common resonance in a system of coupled circuits. It is much easier to tune the two circuits in resonance separately and then to choose an optimum coupling. In this case only three operations are required. Such a tuning procedure provides the so called *full resonance* in a system but requires three tuning organs, which, of course, is a disadvantage.

In cases when a sufficiently wide passband with a reasonable suppression of frequencies outside of it has to be provided by means of tuning, this can be done by choosing the most suitable coupling between the circuits. Sometimes this can be achieved by mutual detuning the two circuits.

Chapter IV. ELECTRICAL FILTERS

4-1. PURPOSE OF FILTERS

In most radio circuits, currents of quite different frequencies flow simultaneously: from very high radio frequencies to low (audio) frequencies and even currents of zero frequency (i.e. direct currents). The currents of certain of these frequencies are used to excite the succeeding elements of an electric circuit, while the currents of other frequencies are considered unwanted as they prevent normal operation of the equipment. Thus, it is necessary to separate currents of certain frequencies from those of all other frequencies. This can be accomplished with the help of special devices known as *electrical filters*.

Filters are divided into classes, according to the frequencies they are intended to transmit to the following circuit: *low-pass* and *high-pass* filters. In addition, *bandpass* filters are also employed whose function is to transmit only currents whose frequencies lie within a desired band ranging from f_1 to f_2 , as well as *band-elimination* filters that attenuate currents whose frequencies lie within a given band. Low-pass filters are subdivided into two main groups: the filters designed to transmit the currents of low audio frequencies and the filters for dc supply sources, which transmit only direct current.

4-2. FILTERS FOR DC SUPPLIES

Direct current can be separated from alternating currents by means of a capacitor. For instance, if a circuit includes a dc supply which also produces alternating current (Fig. 4-1), but only direct current is allowed to pass through the load, a capacitor C must be connected across the load. If the capacitance C

is high enough, almost all the alternating current will pass through it and, practically, bypass the load, thus only direct current will flow through the load. If almost all the alternating current is to flow through the capacitor, its opposition (reactance) to alternating current must be considerably less than the load resistance, thus, we can write

$$\frac{1}{\omega C} \ll Z_L \quad (4-1)$$

Consequently, the capacitance of the capacitor must be chosen the greater, the smaller the load resistance and vice versa. When the load resistance is increased, one may use capacitors of a relatively smaller capacitance. If several alternating currents of different frequencies flow simultaneously in the circuit considered, the capacitor must be selected in such a way as to transmit the lowest of these frequencies.

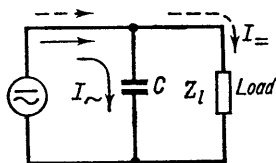


Fig. 4-1. Capacitor as a filter

When a load with a low resistance is used in the circuit or the supply produces an alternating current of very low frequency, the filtering capacitor with a very high capacitance can be required. Such capacitors, however, are bulky and expensive. In these cases it is reasonable to connect in series with the load a choke coil L with an iron core, which possesses a very high inductance (Fig. 4-2). A choke coil wound with a wire of a relatively large diameter offers a low resistance to direct current and a very high reactance to alternating current. Such a filter is called *L-section filter* consisting of a capacitor and an inductance coil. It can be improved by connecting across the load an additional capacitor C (Fig. 4-3) through which the remainder of the alternating current will flow after passing the choke L of the L-section filter. This modified filter is called *Π-section filter*.

More and more complicated filters can be designed by connecting other chokes and capacitors which will provide better filtering of alternating currents. Such filters are composed of several L- or Π-sections connected in series. In practice, however, the presence of more than three sections in a filter does

not improve filtering significantly, but increases considerably the weight, size and cost of the device; therefore, more complicated filters are used very seldom.

If the supply has a low internal resistance, a choke can be used alone as a simple filter (Fig. 4-4). If the ac reactance of the choke is so high that $\omega L \gg Z_l$, then the value of alternating

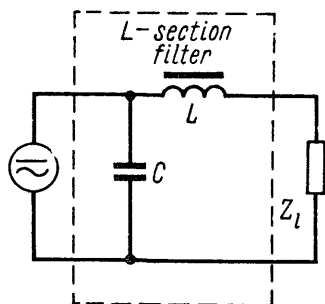


Fig. 4-2. L-section filter

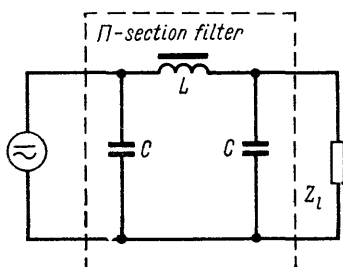


Fig. 4-3. Pi-section filter

current flowing through it will be negligible, whereas the value of direct current to which the choke offers a low resistance can be relatively large. To improve filtration, a capacitor is

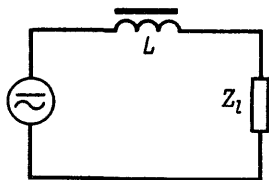


Fig. 4-4. Choke as a filter

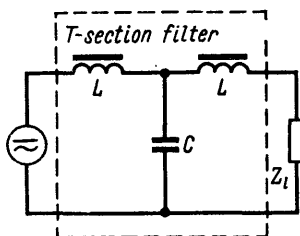


Fig. 4-5. T-section filter

connected in parallel with the load. Such a filter is also called L-section; its filtering action can be further improved by connecting an additional choke into the load circuit. The resulting filter is called T-section filter (Fig. 4-5).

Under load-impedance variations, Pi-section filters provide a more stable output voltage, whereas T-section filters, a

more stable output current. In the case of a T-section filter, the choke connected between the capacitor and the load will prevent large variations of the load current caused by sudden variations of the load impedance. In fact, the emf self-induced in the choke under such variations tends to maintain the same level of the current. When a Π -section filter is used, the voltage across the filter capacitor caused by a sudden change in the load impedance will change quite slowly, while the current flowing through the load will change sharply. Thus, the capacitor will maintain the load voltage at the same level.

The disadvantages of chokes that limit their use in filters are big size, weight and cost. Therefore, resistors are often employed instead of chokes (Fig. 4-6), provided the filters must not meet very rigorous requirements. This helps to reduce considerably the size and cost of filters. The main disadvantage of such filters is that a high power is dissipated in the resistors; therefore, filters of this type are used only in the circuits where relatively small currents are flowing.

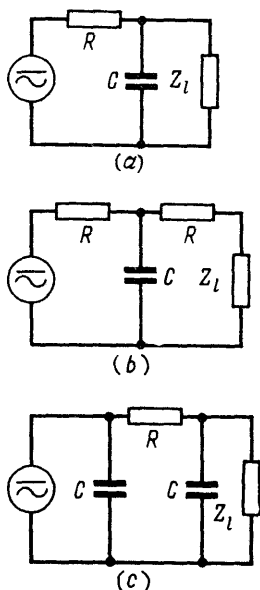


Fig. 4-6. Filters composed of capacitors and resistors
(a) L-section; (b) T-section;
(c) Π -section

4-3. LOW-PASS FILTERS

In radio equipment it is often necessary to transmit alternating currents below a given frequency which is called *cut-off frequency*, $f_{1\text{ cutoff}}$, and to suppress all other currents. For this purpose, low-pass filters are used. They are identical to the filters used for dc supplies except for the values of their parameters such as capacitances, inductances and resistances.

It is impossible to build an ideal filter which will attenuate completely all the currents whose frequency is higher than a

given cut-off frequency. The attenuation of unwanted voltages or currents provided by a filter is estimated by the ratio of their amplitudes at the input of the filter to those at its output. However, the logarithm of this ratio rather than the ratio itself is commonly used. This is due to the fact that the response of the human's organs of vision and hearing, as has been proved experimentally, is proportional to the logarithm of a change in the stimulus rather than to the change itself. Therefore, any variation of attenuation is generally expressed in decibels:

$$\beta = 10 \log_{10} \frac{P_1}{P_2} = 20 \log_{10} \frac{V_1}{V_2} \quad (4-2)$$

Note that the factor of the voltage-ratio logarithm in (4-2) is doubled; this follows from the fact that the power ratio is equal to the ratio of the squared voltages since the powers and voltages are measured across the same terminals.

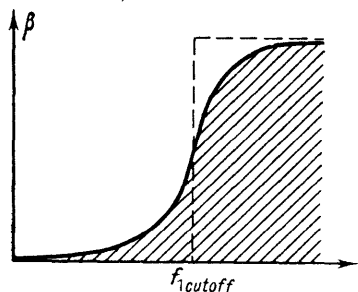


Fig. 4-7. Amplitude-frequency response of a low-pass filter. β is the attenuation factor

The attenuation provided by a low-pass filter is plotted in Fig. 4-7 as a function of frequency of the voltage applied to the filter. To obtain the optimal frequency response of the filter, i.e. the one with the steepest slope at the cut-off frequency, the following requirements should be satisfied: (1)

losses in the filter elements must be minimal, (2) the number of the filter sections must be as great as possible and (3) the load impedance must correspond to the values of the filter parameters.

Let us consider these requirements in more detail. For this purpose we divide any complex filter into elementary L-sections each consisting of a capacitor and a coil, i.e. each having two reactances of opposite sign, X_1 and X_2 (Fig. 4-8). For instance, when a T-section filter is divided into two L-sections (Fig. 4-9), the capacitor C_1 is considered as if it were the parallel combination of two capacitors each having the capacitance $C_1/2$; when a Π -section filter is divided into

two L-sections (Fig. 4-10), the coil L_1 is considered as if it were the series combination of two coils each having the inductance $L_1/2$. The process of transmitting power from one filter section to another can be presented in the following way. The alternating voltage applied to the input of the filter excites a current in its first section, which produces an alternating voltage across the element X_2 , i.e. at the input of the second section, thus, a current is excited in the second section, etc.

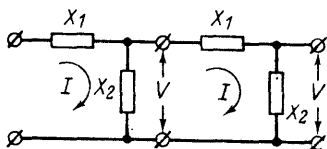


Fig. 4-8. Representation of a filter as a chain of L-sections

If we discount the losses in the filter and assume that all the power supplied to its input is transmitted to the output, then currents and voltages in all sections will be the same. Under these operating conditions which are the most optimal, a loaded filter presents a pure resistance to the supply since all its power is transferred to the filter. This is possible only if the filter has a resistive load and the currents in and the voltages across each filter reactive element develop the same reactive power.

Let us assume that these conditions are satisfied, in other words, the currents in the coils of each elementary section equal I and the voltages across the capacitors of each elementary section equal V . Then, the reactive power developed across each coil is

$$P_{L \text{ react}} = \frac{1}{2} I^2 X_L \quad (4-3)$$

and that across each capacitor,

$$P_{C \text{ react}} = \frac{1}{2} \frac{V^2}{X_C} \quad (4-4)$$

Since $P_{L \text{ react}} = P_{C \text{ react}}$ we can determine the resistance offered by the filter to the supply current:

$$R_{in} = \frac{V}{I} = \sqrt{X_L X_C} \quad (4-5)$$

The same input resistance will be offered by every succeeding filter section to the preceding one. This quantity is

called *wave* or *characteristic impedance* W of the filter. Substituting expressions for reactances $X_L = \omega L_1/2$ and $X_C = 2/\omega C_1$ into equation (4-5), we have

$$R_{in} = W = \sqrt{\frac{L_1}{C_1}} \quad (4-6)$$

For the last section of the filter to operate as the preceding ones, it is necessary to match the load to the filter impedance,

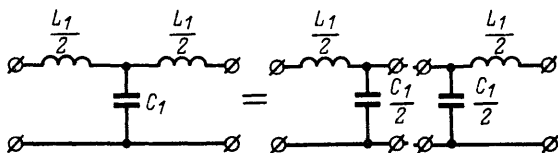


Fig. 4-9. Representation of a T-section low-pass filter as two L-sections

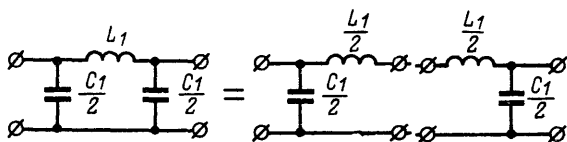


Fig. 4-10. Representation of a Pi-section low-pass filter as two L-sections

i.e. the load must be purely resistive and equal to the filter wave impedance

$$R_{load} = \sqrt{\frac{L_1}{C_1}} = W \quad (4-7)$$

Each L-section is a series oscillatory circuit. At the resonance frequency, the reactance of each capacitor is equal to that of each inductance coil. At lower frequencies the reactance of capacitors rapidly grows, and that of inductance coils falls. Therefore, the current whose frequency is lower than the resonance frequency, for which the coils present a low reactance and the capacitors, a high reactance, is transmitted almost completely through the filter from the source to the load. At the frequencies higher than the resonance frequency, the inductive reactance grows and the capacitive falls. Therefore, the current whose frequency is higher than the re-

sonance frequency, for which coils present a high reactance and capacitors, low reactance, flows through the capacitors and thus does not reach the load.

The resonance frequency of L-sections is the *cutoff frequency of the filter*,

$$f_1 = \frac{1}{2\pi \sqrt{\frac{L_1}{2} \frac{C_1}{2}}} = \frac{1}{\pi \sqrt{L_1 C_1}} \quad (4-8)$$

The parameters of the filter elementary sections determined from expressions (4-7) and (4-8) by the given values of cutoff frequency and load resistance, are

$$L_1 = \frac{R_{load}}{\pi f_1} \quad (4-9)$$

$$C_1 = \frac{1}{\pi f_1 R_{load}} \quad (4-10)$$

Going back to ordinary filters, it is easy to conclude that in Π -section filters all coils must have an inductance L_1 , the capacitors connected at both ends of the filter must have a capacitance $C_1/2$, and those connected between the coils, a capacitance C_1 . In T-section filters all capacitors must have a capacitance C_1 , the coils connected at both ends of the filter must have an inductance $L_1/2$, and those connected between the capacitors, an inductance L_1 .

A change in impedances of filter elements when passing through the cutoff frequency and, consequently, the sharpness of cutoff will be the less, the greater the losses in the filter elements. As the number of sections in a filter is increased, the sharpness of cutoff rises; if, however, this number exceeds 3 or 4, the sharpness of cutoff will rise insignificantly. If the filter output impedance is not matched to that of the load, the filter amplitude-frequency response greatly deteriorates; this can be seen from the experimentally measured characteristics of a matched and mismatched filter shown in Fig. 4-11.

The choice of T-section or Π -section filter is usually determined by specific conditions. For instance, if currents of a low-frequency Ω and a high-frequency ω are flowing in the line *A* (Fig. 4-12) and it is necessary to pass the low-frequency current through the line *B* without preventing the flow of the high-frequency current through the line *A*, only a T-se-

ction filter can be used because, if a Π -section filter is employed, the capacitor connected at its input will present a short circuit for high-frequency currents.

4-4. HIGH-PASS FILTERS

Filters of this type must transmit currents of all frequencies above a given frequency, which is also called cutoff fre-

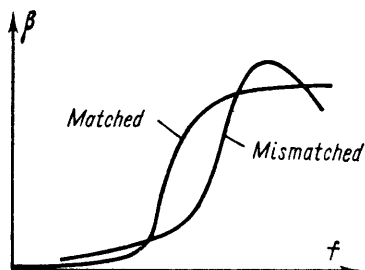


Fig. 4-11. Amplitude-frequency responses of a matched and a mismatched low-pass filters

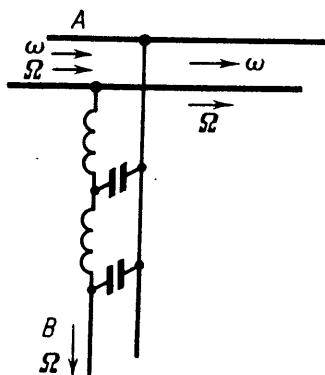


Fig. 4-12. Connection of a T-section filter into a line

quency, and substantially attenuate currents of all frequencies lower than this frequency. Such a filter can be obtained from a low-pass filter (performing the opposite function)

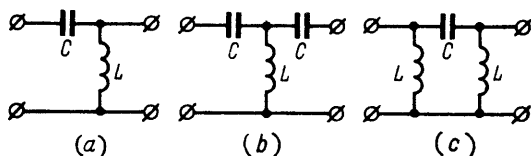


Fig. 4-13. Simple high-pass filters
(a) L-section; (b) T-section; (c) Π -section

by replacing the elements which strongly attenuate high-frequency currents with the elements well transmitting these currents, and vice versa. In other words, the coils in the low-

pass filter must be replaced with capacitors and the capacitors with coils (Fig. 4-13).

The typical amplitude-frequency response of a high-pass filter is given in Fig. 4-14. As in the case of low-pass filters, this response the closer approaches the ideal, the smaller are the losses in the filter, the greater is the number of sections in the filter and the better it is matched to the load. The wave impedance of a high-pass filter is again determined from the equality of reactive powers developed across reactive elements of opposite sign which form L-sections of any filter. For instance, both a T-section filter (Fig. 4-15) and a Π -section filter (Fig. 4-16) can be considered as composed of two L-sections, if we assume that the chokes L_1 are composed of two parallel-connected chokes, each with an inductance $2L_1$, and the capacitors C_1 of two series-connected capacitors, each with a capacity $2C_1$.

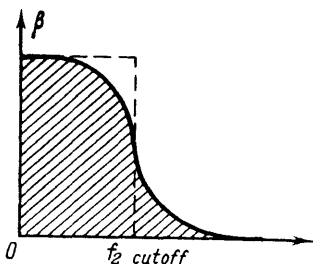


Fig. 4-14. Amplitude-frequency response of a high-pass filter

Using (4-5), we determine the filter wave impedance. As $X_1 = 1/2 \omega C_1$ and $X_2 = 2\omega L_1$, the wave impedance of a high-pass filter is also given by the formula

$$W = \sqrt{\frac{L_1}{C_1}}$$

Each section of the filter presents an oscillatory circuit whose resonance frequency is

$$f_2 = \frac{1}{2\pi \sqrt{2L_1 2C_1}} = \frac{1}{4\pi \sqrt{L_1 C_1}} \quad (4-11)$$

At the resonance frequency the reactances of filter section elements are equal. As the frequency grows, the reactance of capacitors rapidly falls, whereas that of coils increases. Therefore, the current whose frequency is considerably higher than the resonance frequency will pass from the source through the filter capacitors to the load without being appreciably shunted by the coils. The equivalent impedance

of any filter section at such frequencies will be of the capacitive type. Consequently, the filter will transmit the frequencies higher than f_2 . At lower frequencies capacitive reactance

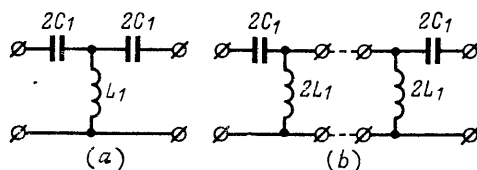


Fig. 4-15. Representation of a T-section high-pass filter (a) as two L-sections (b)

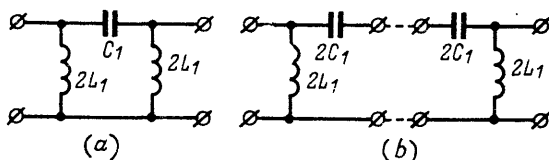


Fig. 4-16. Representation of a Π-section high-pass filter (a) as two L-sections (b)

will grow, and inductive reactance will drop, so that the current supplied by the source will be shunted by the coils without reaching the load. The greater the number of sections in the filter and the smaller the losses in them, the stronger will be the shunting effect. Consequently, the section resonance frequency f_2 is the filter cutoff frequency.

If the cutoff frequency and load resistance are given, by using the condition $R_{load} = W$, one can easily obtain the formulas for calculating the parameters of a high-pass filter

$$L_1 = \frac{R_{load}}{4\pi f_2} \quad (4-12)$$

$$C_1 = \frac{1}{4\pi f_2 R_{load}} \quad (4-13)$$

Considering now a low-pass filter as a whole, we can easily conclude that in T-section filters the capacitors connected at both ends of the filter must have a capacitance $2C_1$, the capacitors connected between the coils, a capacitance C_1 ,

and all coils must have an inductance L_1 . In Π -section filters the coils connected at both ends must have an inductance $2L_1$, the coils connected between the capacitors, an inductance L_1 ; and all capacitors must have a capacitance C_1 .

4-5. BANDPASS AND BAND-ELIMINATION FILTERS

In radio equipment it is often necessary to transmit currents of a given frequency band ranging from f_1 to f_2 . This can be achieved by connecting in series a low-pass filter with cutoff frequency f_2 and a high-pass filter with cutoff frequency f_1 . The frequency characteristic of such a filter is given in Fig. 4-17.

The bandpass filters of this type can be combined both from T-sections (Fig. 4-18) and Π -sections (Fig. 4-19). In practice, however, oscillatory circuits (Fig. 4-20) are used as the elements of bandpass filters.

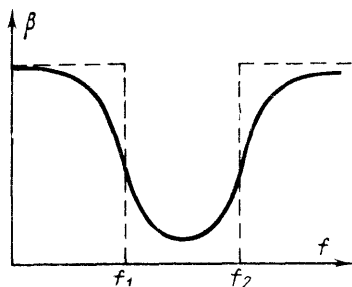


Fig. 4-17. Amplitude-frequency response of a bandpass filter composed of high- and low-pass filters

A series and a parallel circuits offer a low and a very high resistances, respectively, to currents whose frequencies are close to the resonance frequency. Therefore, these currents

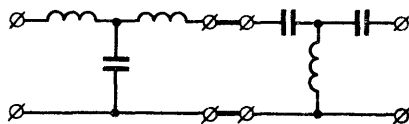


Fig. 4-18. T-section bandpass filter

are transmitted through the filter almost without attenuation. For currents whose frequencies differ considerably from the resonance frequency, a series and parallel circuit presents a high and a very low resistances, respectively, so that these currents are not transmitted through the filter.

A single circuit, however, cannot provide sufficiently uniform characteristics within a wide frequency band. There-

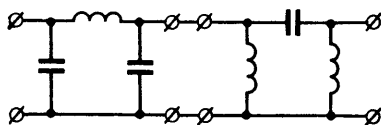


Fig. 4-19. II-section bandpass filter

fore, when a filter which can transmit uniformly currents within a very wide frequency band is required, a system of

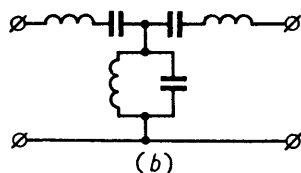
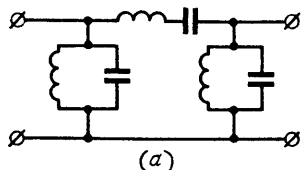


Fig. 4-20. Bandpass filters composed of oscillatory circuits

(a) II-section; (b) T-section

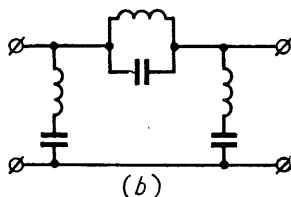
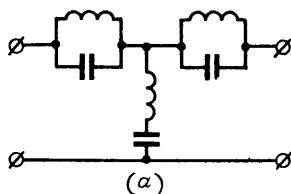


Fig. 4-21. Band-elimination filters

(a) T-section; (b) II-section

several coupled circuits is used. In this case a wide filter bandwidth and uniform transmission of the currents of desired frequencies are achieved either through the use of a tight coupling between the coupled circuits or by detuning them relatively to each other provided the number of coupled circuits in the filter is great enough.

Often a filter must perform a directly opposite function, i.e., it must not transmit currents within a given frequency band. It is clear that such a band-elimination filter can be

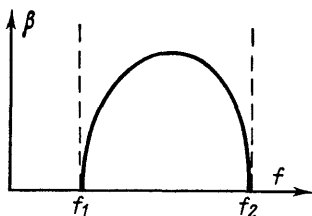


Fig. 4-22. Amplitude-frequency response of a band-elimination filter

obtained from a bandpass filter simply by interchanging the position of the filter elements connected in series with the source and those connected in parallel (Fig. 4-21). The amplitude-frequency response of a band-elimination filter is shown in Fig. 4-22.

5-1. ELECTRICAL SIGNALS IN IDEAL INFINITELY LONG LINES

The preceding chapters described free and forced oscillations in devices of limited size. Many real radio systems, however, have a very big size. To understand what is meant

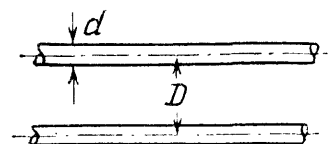


Fig. 5-1. Balanced two-wire line

by "a very big size", we examine the phenomena which take place in an ideal system (i.e. where no losses occur) consisting of two identical straight conductors of unlimited length; such a system is called *ideal balanced two-wire line* (Fig. 5-1).

The conductors carry no current until the line is connected to a supply source. When a dc voltage source E (Fig. 5-2) is connected to the input of the line, current starts to flow in the sections of the line conductors which are close to the source, since the electrons from the source negative terminal enter one of the conductors, and those in the other conductor move to the positive terminal. As a result, the conductor sections close to the source will be charged, and a voltage equal to that of the source will appear across the conductors; thus, the corresponding conductor section will act as a supply source for the succeeding sections, etc.

Consequently, the current will flow along such an infinitely long line from one of its sections to another, setting up a magnetic field around the wires. It is said that an *electromagnetic perturbation* propagates along the line, its speed v being very high but still finite, close to the speed of light.

This, however, does not mean that the electrons which have been injected at a given moment from the source will appear in a second at a distance of 300 000 km away from it. The speed of electron flow is thousands of times less, but the excess of electrons at the beginning of the line produces electron displacement along the whole line, like a displacement of links in one and the same chain. In the case of a

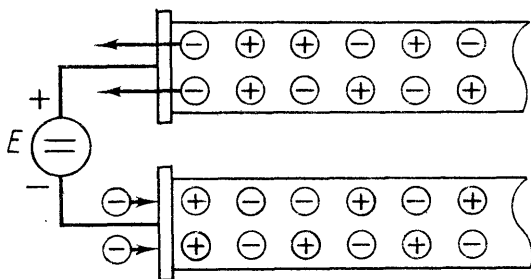


Fig. 5-2. Processes in a line when a dc voltage source is connected to its input

chain, it is necessary to distinguish the speed of movement of separate links, which may be very slow, from the speed of transmission of a push along the chain, which can be tremendous. The role of such "links" in a transmission line is played by electrons, and the push (i.e. signal) being transmitted is the change of electromagnetic field, which sets into motion all the electrons in the line. Therefore, the primary cause of electron flow is always the variations in the field, which propagate at a tremendous speed.

Analogous phenomena occur in a transmission line when a source of alternating emf is connected to its input. If the voltage across the terminals of a generator which is connected to the line is maximum (Fig. 5-3), then beginning with the moment $t=0$ (Fig. 5-3a), the voltage will propagate along the line and at some moment $t=t_1$ (Fig. 5-3b) it will reach a point spaced at $x_1=vt_1$ from the line input. During this time interval the voltage across the generator terminals will drop to a value u_1 . The points lying between the line input and the point x_1 will have intermediate values of the generator voltage. After the time interval $t_2=T/4$ (where T is the gene-

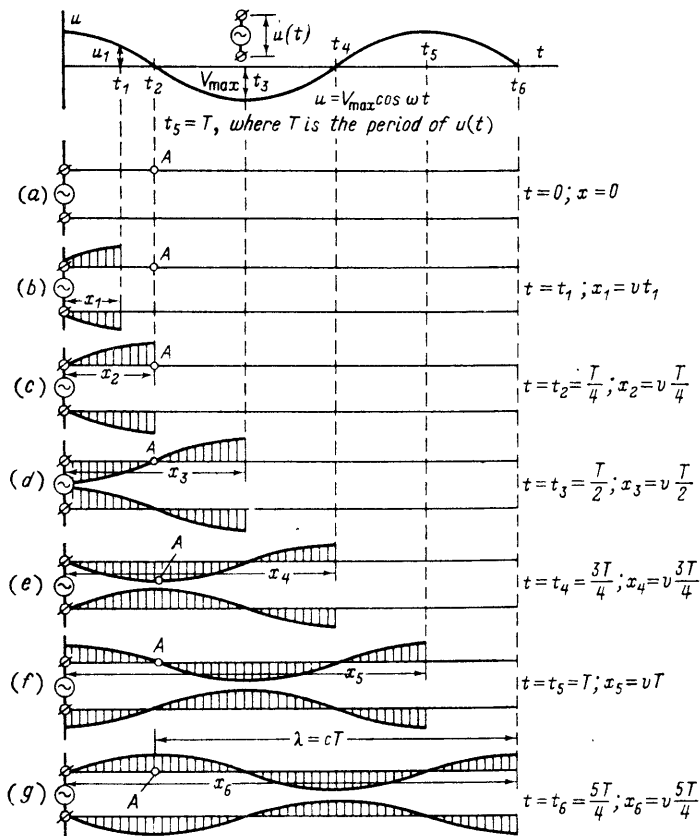


Fig. 5-3. Phenomena in an infinitely long line when an alternating emf source is connected to its input

rator voltage period) the maximum voltage will reach a point spaced at a distance $x_2=vT/4$ from the line input, and the voltage across the generator terminals will drop to zero (Fig. 5-3c). In the next quarter of the period, at time $t_3=$
 $=T/2$ (Fig. 5-3d), the generator voltage will again become maximum but with the negative sign. By this time the original (maximum) voltage will reach the point $x_3=vT/2$, and the zero voltage, the point x_2 . In the same way we can

trace this process for the succeeding moments of time $t_4 = 3T/4$, $t_5 = T$ and $t_6 = 5T/4$.

Thus, alternating voltage propagates along the line producing a varying electric field (Fig. 5-4a). This alternating voltage sets up an alternating current in the conductors,

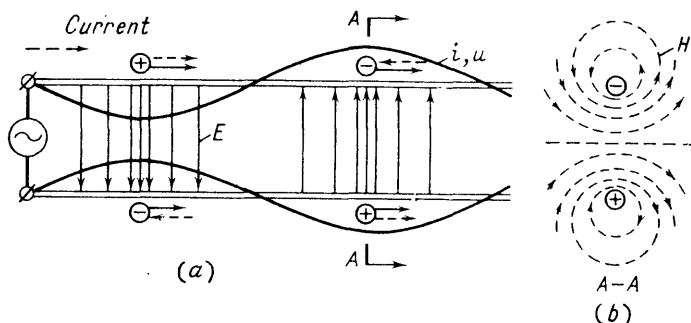


Fig. 5-4. Electric and magnetic fields in a two-wire line

which, in turn, generates a varying magnetic field around the conductors (Fig. 5-4b).

When an alternating emf source is connected to an infinitely long line the so-called *travelling electromagnetic waves* propagate along the line in one direction. The waves of alternating current and voltage propagate along the line simultaneously. As the waves propagate along the line (Fig. 5-3), the voltage, current, and intensity of the electric and magnetic fields assume the same values at certain distances from each other. The distance between two consecutive points where an electromagnetic wave produces the field of identical strength is called *wavelength*.

If at a point A (Fig. 5-3) the alternating voltage has the positive maximum value at the moment t_2 , then the same voltage value will be found at this point after a time interval equal to the period of the alternating voltage. During this time the preceding maximum will move over the distance

$$\lambda = vT \quad (5-1)$$

As the period of alternating voltage is the reciprocal of the frequency ($T=1/f$), the wavelength can be calculated

from the formulas

$$\lambda = \frac{v}{f} \quad (5-2)$$

$$\lambda = 2\pi \frac{v}{\omega} \quad (5-3)$$

The voltage wave produced by the generator arrives at different points of the conductor with a time lag (or delay) depending on the distance of these points from the line input (Fig. 5-3). At the moment $t=0$, when the generator voltage is maximum, there is no voltage at the point A . The maximum of the voltage wave will reach the point A in a time $t_2 = T/4$; by this moment the generator voltage will have dropped to zero. In other words, if the generator voltage follows the law $u = V \cos \omega t$, it will vary at the point spaced at a distance x from the line input in a similar way, but with a time lag equal to the time interval $\Delta t = x/v$ required for the electromagnetic wave to reach this point. Consequently, the voltage at this point varies by the law

$$u_x = V_{\max} \cos \omega \left(t - \frac{x}{v} \right) \quad (5-4)$$

The expression (5-4) is called travelling wave equation. With its help it is possible to determine voltage distribution along the line for a fixed moment of time t or to find the voltage at a fixed point x as a function of time.

From the above discussion it is clear that the phenomena which occur in transmission lines are quite different from those taking place in coils, capacitors or resistors. The latter elements are characterized by the fact that the phases and amplitudes of electric oscillations are the same at all their points. In the lines, however, whose length is comparable with the wavelength of a signal propagating along the line, the signal phase is different at different points. Consequently, a transmission line can be called "long" if its length is comparable with the signal wavelength. In practice the time required for a signal to reach the opposite end of the line, i.e. the time delay, is usually taken into account when the line length becomes more than 0.1λ . Moreover, in the systems described earlier one can easily point out the elements where the electromagnetic energy is stored (coils and capacitors),

whereas in the transmission lines both the electric and magnetic energy is stored along the entire conductor.

The circuits examined earlier are called *lumped element systems*. A transmission line, on the other hand, is a typical

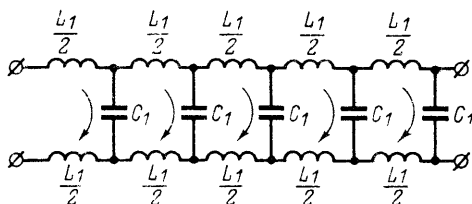


Fig. 5-5. Equivalent circuit of a transmission line

example of a *distributed element system*. In fact, any conductor of unit length has a certain inductance. Also, two adjacent conductors of a transmission line section, which are separated by a dielectric material form a capacitor which has a certain capacitance. Thus, any transmission line is characterized by an *inductance* and a *capacitance per unit length of line*. Imagine that the whole line is divided into elementary sections, then its equivalent electrical circuit can be represented as a series-connected chain of cells

consisting of inductances L_1 and capacitances C_1 interconnected by wires (Fig. 5-5). Such an equivalent circuit enables us to explain a number of important properties associated with transmission line operation.

When an electromagnetic wave propagates along the line, the electric and magnetic energy stored in its separate sections varies. If, during a short time period Δt , the wave has moved along the line by distance Δx (Fig. 5-6) which is so small that the current can be considered to be the same along the entire section Δx , the increment of the magnetic flux in this section caused by the incoming wave will be equal to the section inductance multiplied by the current increment in it. In the case considered the magnetic flux increment is

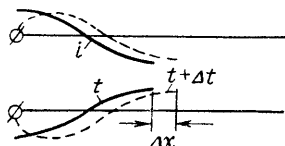


Fig. 5-6. Variation of current distribution along the line for a short time period

equal to the current flowing in this section, as there was no current prior to the arrival of the wave, i.e. $\Delta\Phi = Li = L_1 i \Delta x$. This change in the flux produces a self-induced emf in the conductor section, which is

$$e = \frac{\Delta\Phi}{\Delta t} = L_1 i \frac{\Delta x}{\Delta t} \quad (5-5)$$

(the signs of the emf will not be considered as they are opposite in the two conductors). Under the action of this emf a charge $\Delta q = C_1 e \Delta x$ is accumulated in the distributed capacitance of the section; this charge is carried by the current i during the time period Δt , i.e. $\Delta q = i \Delta t$. From these two equations, we have

$$e = \frac{i}{C_1} \frac{\Delta t}{\Delta x} \quad (5-6)$$

Equating expressions (5-5) and (5-6), we obtain $(\Delta x / \Delta t)^2 = 1 / L_1 C_1$; but $\Delta x / \Delta t$ is the velocity of an electromagnetic wave propagating along the line. Consequently,

$$v = \frac{1}{\sqrt{L_1 C_1}} \quad (5-7)$$

The velocity of an electromagnetic wave propagating along the line depends on the values of capacitance and inductance per unit length of line. The greater they are, the less the velocity of wave propagation. As can be shown, the capacitance (in farads) per metre of a balanced two-conductor line can be found from the formula

$$C_1 = \frac{\epsilon}{4 \ln \frac{2D}{d}} \frac{1}{9 \times 10^9} \quad (5-8)$$

and the inductance (in henries) per metre, from the formula

$$L_1 = 4\mu \ln \frac{2D}{d} 10^{-7} \quad (5-9)$$

where D and d are the distance between the axes of the two conductors and the conductor diameter, respectively (see Fig. 5-1); ϵ and μ are the dielectric constant and magnetic permeability of the medium surrounding the conductors, respectively.

If we substitute expressions (5-8) and (5-9) into (5-7), the velocity of propagation of an electromagnetic wave (in metres per second) equals

$$v = \frac{3 \times 10^8}{\sqrt{\epsilon\mu}} \quad (5-10)$$

In free space $\epsilon = \mu = 1$, and the velocity of propagation is equal to 3×10^8 m/s. The velocity of an electromagnetic wave propagating along conductors is the same as in free space, and is determined by the dielectric constant ϵ and magnetic permeability μ of the medium surrounding the conductors. For conductors separated by an air gap, $\epsilon \approx \mu \approx 1$ and $v \approx 3 \times 10^8$ m/s.

As a wave travels along the line, the latter absorbs all the energy supplied by the generator, returning nothing; therefore, it presents a purely resistive load for the generator. Consequently, the current and voltage in the line along which the wave propagates must be in phase. In each section of the transmission line magnetic field energy is stored in the distributed inductance $W_L = (L_1 \Delta x I_{\max}^2)/2$, where $I_{\max}$ is the amplitude of the line current. At the same time electric field energy is stored in the distributed capacitance $W_C = (C_1 \Delta x V_{\max}^2)/2$, where $V_{\max}$ is the amplitude of the line voltage. The energy stored by the inductance must be equal to that stored by the capacitance, otherwise the line would present a certain inductive or capacitive load for the generator. Thus, $(L_1 \Delta x I_{\max}^2)/2 = (C_1 \Delta x V_{\max}^2)/2$, hence

$$\frac{V_{\max}}{I_{\max}} = \sqrt{\frac{L_1}{C_1}} = Z_0 \quad (5-11)$$

Here Z_0 is the line input impedance for the travelling wave, called the *wave* or *characteristic impedance* (it is a pure resistance).

In the case of a travelling wave the line input impedance is identical at all points of the line. In fact, if we take any point along the line, then to the right of this point we shall always see an infinitely long line whose input impedance is equal to its wave impedance.

The distributed capacitance and inductance of the line depend only on the diameter of its conductors and the distance between them. Consequently, its wave impedance is

also fully determined by the design parameters of the line and does not depend on the generator frequency. Substituting expressions (5-8) and (5-9) into (5-11), we have

$$Z_0 = 276 \log_{10} \frac{2D}{d} \quad (5-12)$$

The distance D between the axes of the two conductors and the conductor diameter d are selected from design considerations to obtain small losses, high mechanical strength, etc. For two-conductor aerial lines, the wave impedance usually ranges from 300 to 800 Ω .

5-2. SIGNALS IN FINITE LENGTH LINES WITH FAR END OPEN-CIRCUITED

A transmission line of finite length with far end open-circuited offers an infinitely great resistance to direct or low-frequency currents, when it is under steady-state condition. For high-frequency currents, however, for which the size of the line becomes comparable with the operating wavelength, the line impedance may be finite or even zero.

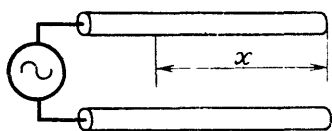


Fig. 5-7. Section of an open-circuited line

When examining the phenomena which occur in the line of finite length (Fig. 5-7), it is convenient to measure the

distances from its far end. When the line is connected to a source, an electromagnetic wave starts to propagate along it; this wave is called incident wave. As all the changes in voltage applied to the line input are transmitted along the line, after some time the voltage across the far end of the line produced by the incident wave will vary by the law

$$u_{x=0} = V_{\max} \cos \omega t \quad (5-13)$$

Since the electrons reaching the far end of the line cannot move farther, the current in the far end section of the line is

$$i_{x=0} = 0 \quad (5-14)$$

in spite of the presence of voltage at the far end.

The same conditions would take place at the far end of the line if it were connected to a generator operating in opposition to the incident wave and producing an emf whose amplitude and phase were the same as for the incident wave. In fact, two identical voltages would then act in opposite directions on the electrons at the far end of the line, so that there would be no current at the point $x=0$ in spite of the presence of incoming voltage waves.

As all variations which occur at the line far end are transmitted along the line at the speed of light, the entire line will operate in such a way as if the generator were connected at its end, producing the above-mentioned emf. This generator produces a wave travelling from the far end of the line to its input and having at the line input the same amplitude and phase as the incident wave. Such a wave is called *reflected wave*.

Since the voltage produced by the incident wave at the line far end is determined by expression (5-13), the voltage at the point x at the time t will be produced by the wave which has left the generator after the time period $\Delta t = x/c$. Consequently, at the point x the incident wave will produce the voltage

$$u_{x\ inc} = V_{\max} \cos \omega \left(t + \frac{x}{c} \right) \quad (5-15)$$

The wave reflected from the line far end also arrives at this point but at the instant which precedes the moment t by the time required for the reflected wave to travel the distance x . It will produce at this point the voltage

$$u_{x\ refl} = V_{\max} \cos \omega \left(t - \frac{x}{c} \right) \quad (5-16)$$

The resultant voltage at this point is equal to the sum of the incident and reflected wave voltages:

$$u_x = V_{\max} \left[\cos \omega \left(t + \frac{x}{c} \right) + \cos \omega \left(t - \frac{x}{c} \right) \right]$$

Using a well-known trigonometrical formula

$$\cos \alpha + \cos \beta = 2 \cos \frac{\alpha + \beta}{2} \cos \frac{\alpha - \beta}{2}$$

we obtain

$$u_x = 2V_{\max} \cos \frac{\omega}{c} x \cos \omega t \quad (5-17)$$

As $\omega = 2\pi/T$ and $T = \lambda/c$, the ratio $\omega/c = 2\pi/Tc = 2\pi/\lambda$. Hence,

$$u_x = 2V_{\max} \cos \frac{2\pi}{\lambda} x \cos \omega t \quad (5-18)$$

This expression is called equation of a *standing voltage wave*.

The time-independent factor standing in front of the term $\cos \omega t$ determines the amplitude of the oscillatory process which takes place in the line. In the case considered the voltage amplitude varies along the line. At the far end of the

line, where $x=0$, the voltage $V_{x=0} = 2V_{\max}$, i.e. the voltage is twice as great as that produced by the generator, since the incident and reflected waves at the far end are added. The points where the resultant amplitude reaches its maximum are called *wave loops*.

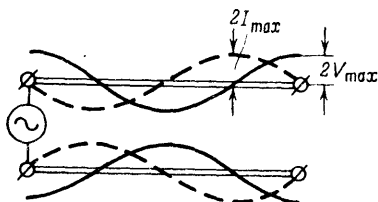


Fig. 5-8. Distribution of current and voltage along an open-circuited line

At a distance $x = 0.25\lambda$ from the line far end the vol-

tage $V_{x=0.25\lambda} = 0$, i.e. at any instant the voltage at this point will be zero due to mutual cancellation of the incident and reflected waves. The points where the resultant amplitude is minimum are called *wave nodes*. At the point spaced $x = 0.5\lambda$ from the line far end the resultant amplitude is again equal to $2V_{\max}$, thus, the voltage takes the same magnitudes at half wavelength intervals (Fig. 5-8). In the traveling-wave mode the voltage amplitude is identical along the entire line whereas in the standing-wave mode it varies from one point of the line to another ranging from double amplitude to zero and assuming the same values in a period equal to half of the wavelength.

It should be noted that Fig. 5-8 gives the amplitude of voltage variations at separate points along the line rather than instantaneous values of the voltage at these points, which vary as shown in Fig. 5-9, where the distribution of

instantaneous values of the voltage along one of the line conductors is plotted for the moments of time t_1, t_2, t_3 , etc.

Similarly, we can obtain the current distribution in the conductors of line with far end open-circuited. As the incident and reflected waves are both travelling waves, the cor-

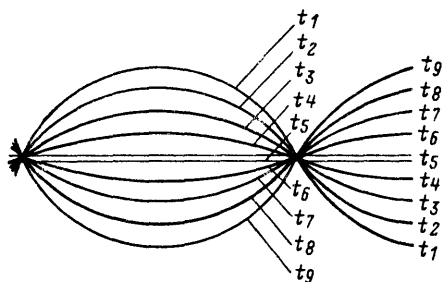


Fig. 5-9. Distribution of instantaneous voltage values in an open-circuited line for various time moments

responding currents are in phase with the voltages. Therefore, on the basis of the expressions (5-15) and (5-16), we can write

$$\left. \begin{aligned} i_{x \text{ inc}} &= I_{\max} \cos \omega \left(t + \frac{x}{c} \right) \\ i_{x \text{ refl}} &= I_{\max} \cos \omega \left(t - \frac{x}{c} \right) \end{aligned} \right\} \quad (5-19)$$

It is only necessary to consider that the reflected voltage wave produces a current which flows from the far end of the line to its input. Therefore, $i_{x \text{ inc}}$ and $i_{x \text{ refl}}$ flow in opposite directions and hence

$$i_x = i_{x \text{ inc}} - i_{x \text{ refl}} \quad (5-20)$$

After some transformations we obtain

$$i_x = -2I_{\max} \sin \frac{2\pi}{\lambda} x \sin \omega t = 2I_{\max} \sin \frac{2\pi}{\lambda} x \cos (\omega t + 90^\circ) \quad (5-21)$$

From this expression it can be seen that the current amplitude also varies along the line, but the character of this variation is different from that of the voltage. In fact, at the far end of the line ($x=0$) the current is always zero; at a dis-

tance $x=0.25 \lambda$ from the far end the current amplitude is equal to the double value of the generator current, etc. The distribution of current amplitudes along the line is shown by the dashed line in Fig. 5-8. It is easy to see that it is displaced relative to the distribution of the voltage amplitudes by a quarter of the wavelength. From expressions (5-18) and (5-21) it follows that the current leads the voltage by 90° . This indicates that the line presents a reactive impedance to the source, which is quite natural since the line does not absorb energy.

If only the travelling wave is present, the line input impedance is the same at any point of the line and is resistive in character; if a standing wave is present, however, the line input impedance is of a reactive type and has different values. In fact, from Fig. 5-8 it can be seen that the line input impedance which is equal to the ratio of the voltage amplitude to the current amplitude is different at different points along the line. For instance, at a distance of 0.25λ from the line far end the current amplitude is equal to $2I_{\max}$ and the voltage amplitude is zero. Consequently, the line input impedance at this point is also equal to zero. At a distance of 0.5λ from the line far end the current is zero, and the voltage is equal to $2V_{\max}$; consequently, the line input impedance at this point is equal to infinity. (If voltage is applied to some element and does not produce current in it, this means that the element has an infinitely great impedance.) The above values of the line input impedance are repeated with a period equal to half of the wavelength. At the intermediate points, the line input impedance can assume any values from zero to infinity.

Inasmuch as a lossless line open-circuited at its far end does not absorb energy, but only stores it during some part of the period and then returns it back to the source, it presents a purely reactive load to the generator. Taking the ratio of the voltage amplitude determined by (5-18) to the current amplitude given by (5-21) and indicating the 90° -phase displacement of the current relative to the voltage by $-j$, we obtain the following formula for the input impedance of the line open-circuited at its far end

$$X_{in} = -j \frac{V_{\max}}{I_{\max}} \cot \frac{2\pi}{\lambda} x \quad .$$

It should be recalled that the ratio of voltage and current amplitudes is the line wave impedance, whence

$$X_{in} = -jZ_0 \cot \frac{2\pi}{\lambda} x \quad (5-22)$$

The input impedance of the lossless line is plotted in Fig. 5-10 as a function of its length. From the curves shown it can be seen that the line input impedance assumes values ranging from $+\infty$ to $-\infty$; in other words, by selecting the line of a corresponding length one can obtain the equivalent capacitance or inductance of any required value. At the points along the line where its reactance becomes equal to zero or infinity the line behaves like a lossless oscillatory circuit at series or parallel resonance, respectively.

As compared with an ordinary oscillatory circuit, transmission lines possess a number of important features. Their input reactance becomes equal to zero or infinity not at one frequency, as in a conventional oscillatory circuit, but at an infinite series of frequencies for which the line length is equal to integer number of $\lambda/4$, i.e. the line has an infinite number of resonance frequencies.

By changing the length of a transmission line it is possible to change its resonance frequency, i.e. to tune the line to one frequency or another. Such a line is often called tunable. When a standing wave is produced in the transmission line, the current and voltage are distributed nonuniformly along the line, which results in a nonuniform distribution of electric- and magnetic-field energy along the line. In the case of a transmission line open-circuited at its far end, the line section near the line far end stores more electric energy than magnetic one since the voltage across the conductors of this

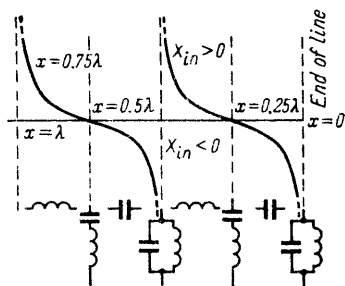


Fig. 5-10. Input impedance of an open-circuited lossless line as a function of its length

section is high while the current in them is low. On the other hand, a line section within which the current is close to its maximum value stores more magnetic energy than electric one. If the length of the line in which a standing wave is produced is equal to an integer number of $\lambda/4$, the amount of magnetic energy stored in this line is equal to that of electric energy, and the line in this case presents a pure resistance to the source. If such a line is made somewhat shorter or longer, its input impedance becomes reactive, character of the latter being determined by the dominant type of energy stored in the line section which has been shortened or lengthened.

5-3. SIGNALS IN FINITE LENGTH LINES WITH FAR END SHORT-CIRCUITED

In short-circuited lines, as in those with the far end open-circuited, the operation mode is determined by the far-end conditions. Since the conductors at the far end of the line are short-circuited, the potential difference between them at this point is zero, i.e. $V_{x=0}=0$. The same conditions would take place at the far end of the line if, instead of a short circuiting jumper, a generator producing a voltage which is equal in amplitude and shifted in phase by 180° relative to that of the incident wave were connected at the line far end. Such a generator would produce a reflected wave shifted at the line far end by 180° relative to the incident wave. Therefore, the distribution of the voltage along the line is again determined by expression (5-16) but in this case with the sign minus in front of its right-hand side.

The phase of the current travelling in the direction opposite to the incident wave is also shifted by 180° , i.e. the direction of the current is changed to the opposite. This means that the currents of incident and reflected waves coincide in phase, and the sign 'minus' in the expression (5-20) must be replaced by 'plus'. In other words, in case of a short-circuited line the current distribution is the same as the voltage distribution along the open-circuited line, and the voltage distribution is the same as was the current distribution. Therefore, without repeating the conclusions of the preceding

paragraph, we may write

$$\begin{aligned} u_x &= -2V_{\max} \sin \frac{2\pi}{\lambda} x \sin \omega t = \\ &= 2V_{\max} \sin \frac{2\pi}{\lambda} x \cos (\omega t + 90^\circ) \end{aligned} \quad (5-23)$$

$$i_x = 2I_{\max} \cos \frac{2\pi}{\lambda} x \cos \omega t \quad (5-24)$$

The 90° -phase displacement of the current relative to the voltage (see Fig. 5-11) indicates that the line input impedance is reactive; in this case, however, as distinct from an open-circuited line, the current lags behind the voltage by 90° , which is marked by the factor $+j$, i.e.

$$X_{in} = jZ_0 \tan \frac{2\pi}{\lambda} x \quad (5-25)$$

The expression (5-25) shows that a short-circuited line can have a reactive input impedance ranging from $+\infty$ to $-\infty$, i.e. it can present a capacitor or a coil with capacitance or inductance depending on the line length expressed in wavelengths (x/λ) which is called *electrical length*. With the length of a short-circuited line being 0.25λ , 0.75λ , 1.25λ , etc., its wave impedance is equal to infinity. In such a case it behaves as an ideal parallel oscillatory circuit tuned in resonance. If at a given operating frequency the length of a short-circuited line is somewhat greater than $n \cdot \lambda/4$ where n is any odd integer then the line offers a capacitive reactance to the generator connected to its input, and inductive reactance when the line length is somewhat less than $n \cdot \lambda/4$. A short-circuited line with a length equal to 0.5λ , 1λ , 1.5λ , etc., has a zero input impedance and behaves as an ideal series oscillatory circuit tuned in resonance. Such a line offers a capacitive or inductive reactance to the input generator when the generator frequency is respectively lower or higher than the line resonance frequency.

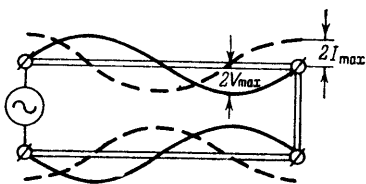


Fig. 5-11. Distribution of current and voltage along a short-circuited line

Figure 5-12 shows the points along the line where the latter is equivalent to a series or parallel circuit, a capacitor or coil. Comparing the curves in Figs. 5-10 and 5-12 one can see that short-circuited and open-circuited lines are interchangeable. In fact, they have identical input impedances when the length of one of them is greater or less than that of the other by $\lambda/4$.

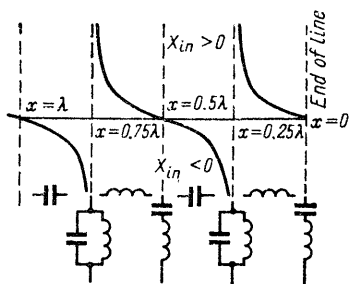


Fig. 5-12. Input impedance of a short-circuited line as a function of its length

5-4. SIGNALS IN LINES WITH A REACTIVE LOAD

If we connect a capacitor or inductance coil at the far end of an open-circuited line, then due to the absence of energy losses, standing waves will be set up near the end of the line and the line input impedance will be a pure reactance. In

this case only the amplitude distribution of the standing wave is changed. Whereas the current at the end of an open-circuited line is always zero, now this no longer will be the case as some current will be flowing through the capacitor or inductance coil connected at the line end.

To explain the character of distribution of standing waves in a line loaded with a reactance, it is convenient to replace the load by the corresponding section of an open-circuited line. Using the curves given in Fig. 5-10, we find the length of the equivalent line section whose reactance is equal to that of the load, and assume that this section is connected at the line far end instead of the load. Then we plot the distribution of currents and voltages for the resultant open-circuited line. The part of the plot corresponding to the original section of the line will represent a new standing wave distribution.

Let us first examine the case when a capacitor is connected at the far end of the line (Fig. 5-13). As the connection of a capacitor is equivalent to the lengthening of an open-circuited line by less than 0.25λ , the first (from the right) voltage minimum (or current maximum) moves closer to the end of

the line, but takes a position before it. Thus, near the far end of the line (moving in the direction of the line beginning), the descending part of the voltage curve and the ascending part of a current curve are located.

If a coil is used as the line load, the length of the equivalent section added to the original line is more than 0.25λ ,

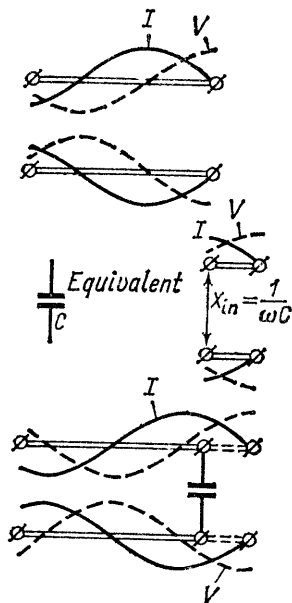


Fig. 5-13. Distribution of current and voltage along a line loaded with a capacitor

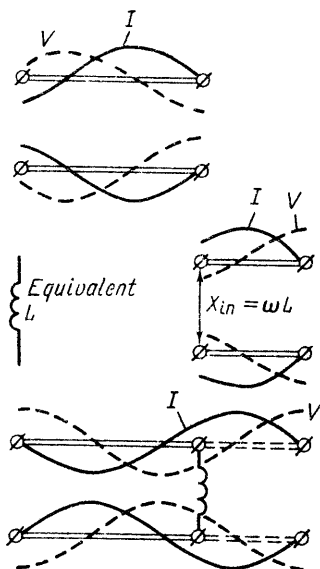


Fig. 5-14. Distribution of voltage and current along a line loaded with an induction coil

and the voltage minimum moves beyond the far end of the original line. Therefore, near the end of the line we shall find the ascending part of the voltage curve and the descending part of the current curve (Fig. 5-14).

If we know the distribution of current and voltage along the line, as well as its wave impedance, we can determine the magnitude and character of the load connected at the line far end.

5-5. SIGNALS IN LINES WITH A RESISTIVE LOAD

If a line is used to transmit high-frequency energy, the load at its far end should be so selected that it absorbs all the energy brought in by an incident wave and produces no reflected wave. To satisfy these conditions the load must be a pure resistance. The load resistance can be determined from an examination of the phenomena which occur in an infinitely

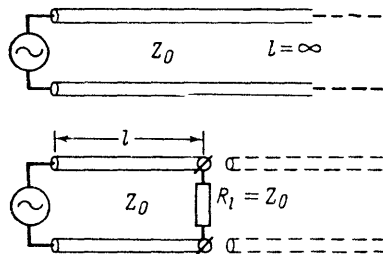


Fig. 5-15. Replacement of an infinitely long line by a pure resistance equal to the line wave impedance

long line, as described in Section 5-1. The input impedance of an infinitely long line is equal to its wave impedance. Therefore, if the line is cut and its infinitely long section is replaced by a pure resistance equal to the line wave impedance (see Fig. 5-15), the phenomena which occur in the remaining section of the infinitely long line will not change, i.e. the travelling waves will propagate along the line and all their energy will be absorbed by the load. Such a line is said to be *matched with the load*. For any other load resistance some part of the energy will not be absorbed by the load and reflected waves will appear in the line.

Let us consider an incident wave and denote the voltage amplitude by V_{inc} ; then the amplitude of the current at all points along the line will be

$$I_{inc} = \frac{V_{inc}}{Z_0} \quad (5-26)$$

At the far end of the line the incident voltage wave produces current

$$I_{load} = \frac{V_{inc}}{R_{load}} \quad (5-27)$$

Such a current would flow in the far end section of the line if it were connected to an ac generator (Fig. 5-16) producing a current which is equal to the current I_{load} less the

current of the incident wave, i.e.

$$I_{refl} = I_{load} - I_{inc} = \frac{V_{inc} (Z_0 - R_{load})}{R_{load} Z_0} \quad (5-28)$$

Such a generator would produce a current and voltage wave which propagates from the far end of the line towards its beginning, i.e. a reflected wave.

The load R_{total} for this generator is the parallel combination of the resistor connected at the line far end and the line input impedance which is equal to the line wave impedance provided there is no reflection from the line input. Consequently,

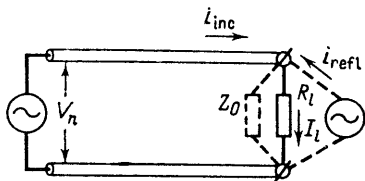


Fig. 5-16. Line loaded with a resistor

$$\frac{1}{R_{total}} = \frac{1}{R_{load}} + \frac{1}{Z_0} \quad (5-29)$$

whence the load resistance for the generator connected at the line far end is

$$R_{total} = \frac{R_{load} Z_0}{R_{load} + Z_0} \quad (5-30)$$

Therefore, the voltage produced by this generator at the line far end, i.e. the reflected wave voltage is

$$V_{refl} = I_{refl} R_{total} = I_{refl} \frac{R_{load} Z_0}{R_{load} + Z_0} \quad (5-31)$$

From expressions (5-28), (5-30) and (5-31) we obtain

$$V_{refl} = V_{inc} \frac{Z_0 - R_{load}}{R_{load} + Z_0}$$

whence the *reflection coefficient* of the line can be defined as

$$p = \frac{V_{refl}}{V_{inc}} = \frac{Z_0 - R_{load}}{R_{load} + Z_0} \quad (5-32)$$

If the load resistance of the line is equal to its wave impedance then the current $I_{load} = I_{inc}$ and the current $I_{refl} = 0$ thus there is no reflection in the line and, according

to formula (5-32), the reflection coefficient is zero. If, however, the load resistance of the line is greater than its wave impedance, a current $I_{load} < I_{inc}$ will flow through the load and, consequently, the current of the equivalent generator connected at the line far end must be subtracted from the incident wave current, i.e. it must have the opposite phase. Therefore, a current node appears at the line far end.

In a reflected wave the voltage and current always have opposite signs. In fact, the positive half-wave of the incident

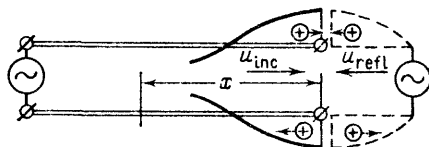


Fig. 5-17. Incident and reflected waves of voltage causing the charges in the conductors to move in opposite directions

wave voltage makes the charges in the upper conductor move towards the load, whereas the positive half-wave of the reflected wave voltage moves them towards the line input, i.e. in the negative direction (Fig. 5-17). Consequently, a voltage loop is produced at the line far end.

If the line load resistance is less than its wave impedance, then $I_{load} > I_{inc}$, thus the current of the equivalent generator at the line far end adds the incident wave current, which happens when their phases coincide. Therefore, in this case the current loop and, correspondingly, the voltage node will appear at the line far end. At quarter-wavelength intervals nodes alternate with loops since the phase of the incident wave at these intervals increases by $\pi/2$, while that of the reflected wave decreases by the same amount (or vice versa). Therefore, the incident and reflected waves will alternatively be added and subtracted at quarter-wavelength intervals.

In the case of a resistive load some part of incident wave energy is always consumed by the load, as distinct from the reactive load. In the former case, therefore, travelling waves which carry the above-mentioned energy are always

present in addition to standing waves. For this reason, the voltages and currents in the line with a resistive load will never reach the double amplitude of the travelling wave at loops and a zero value at nodes; the exceptions are the two extreme cases described above, i.e. a short-circuited line and an open-circuited line for which $R_{load}=0$ and $R_{load}=\infty$, respectively.

The closer the load resistance to the line wave impedance, the smaller the reflection from the line far end, the smaller the amplitude of standing waves and the greater the amplitude of the travelling wave, i.e. the better the line matches the load. This can be seen from a series of line voltage distributions plotted for different load resistances (Fig. 5-18).

The distribution of voltage or current along the line where standing and travelling waves are present is usually characterized by the travelling wave ratio (TWR) which is defined as the ratio of voltage or current at a node to those at a loop, i.e.

$$k = \frac{V_{\min}}{V_{\max}} = \frac{I_{\min}}{I_{\max}}$$

For travelling waves, $k=1$, and for standing waves, $k=0$. The travelling wave ratio characterizes the degree of mismatch

of the line impedance with respect to that of the load. As nodes appear at the points where the reflected wave is subtracted from the incident one while loops appear at the points where the two are added, we have $V_{\max}=V_{inc}+V_{refl}$, while

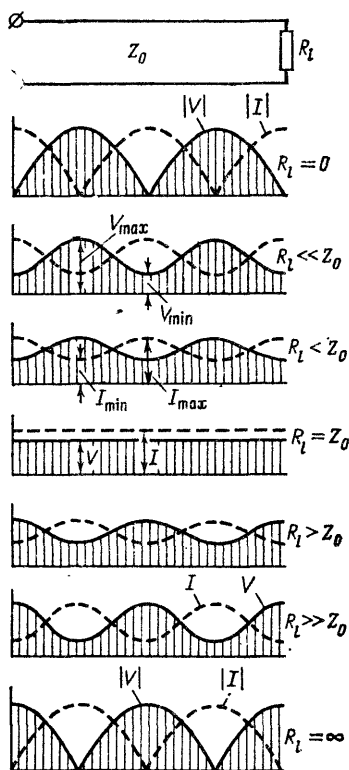


Fig. 5-18. Distribution of current and voltage in a line loaded with different resistors

$V_{\min} = V_{inc} - V_{refl}$. Hence, the travelling wave ratio is

$$k = \frac{V_{inc} - V_{refl}}{V_{inc} + V_{refl}}$$

Dividing the numerator and the denominator by V_{inc} , we obtain

$$k = \frac{1 - \frac{V_{refl}}{V_{inc}}}{1 + \frac{V_{refl}}{V_{inc}}} = \frac{1 - p}{1 + p}$$

Substitution of the expression for the reflection coefficient into this formula gives

$$p = \frac{Z_0 - R_{load}}{Z_0 + R_{load}}$$

Then

$$k = \frac{R_{load}}{Z_0}$$

From (5-32) it can be seen that for $R_{load} > Z_0$ the reflection coefficient becomes negative. This means that the reflected wave phase is reversed (as compared with the case when $Z_0 < R_{load}$), i.e.

$$p = \frac{R_{load} - Z_0}{R_{load} + Z_0}$$

Substituting these equations into the general expression for the travelling wave ratio, when $Z_0 < R_{load}$ we obtain

$$k = \frac{Z_0}{R_{load}}$$

Thus, the travelling wave ratio can be used to determine how many times the load resistance is greater or less than the line wave impedance. Having found experimentally the travelling wave ratio, we can easily answer the question: how many times should we change the load resistance or the line wave impedance in order to match them completely?

In practice, the reciprocal of the travelling wave ratio,

$$\frac{1}{k} = \frac{V_{\max}}{V_{\min}}$$

called the *standing wave ratio*, is often used.

Since the matching conditions require that the line load be a pure resistance equal to the line wave impedance, two lines can be interconnected without producing any reflection at the junction provided their wave impedances are the same. Moreover, at the far end of a line two other lines can be connected also without producing any reflection provided the wave impedance of each line being connected is twice as great as the wave impedance of the main line.

5-6. SIGNALS IN LINES WITH A COMBINED LOAD

In general, when a combined load, i.e. the one consisting of a resistance and a reactance is connected at the line far end, some part of the incident wave energy is absorbed by the load while the remaining part is reflected so that both standing and travelling waves exist simultaneously in the line and its input impedance has both a resistive and reactive component.

In the presence of reflection, the current at an arbitrary point along the line spaced at a distance x from its far end is a sum of the currents of the incident and reflected waves, i.e.

$$i_x = i_{x\ inc} + i_{x\ refl} = I_{\max\ inc} \cos \omega \left(t + \frac{x}{c} \right) + I_{\max\ refl} \cos \left[\omega \left(t - \frac{x}{c} \right) \pm \varphi \right] \quad (5-33)$$

where φ is the angle of an additional phase displacement caused by the reflection of waves from the combined load.

Since only a part of incident wave energy is reflected from the load, the amplitude of the reflected wave is a fraction of that of the incident wave or

$$I_{\max\ refl} = p I_{\max\ inc} \quad (5-34)$$

The factor p (reflection coefficient) expresses the ratio of the amplitude of the reflected wave to that of the incident wave.

Since both the incident and reflected waves are travelling waves, the line presents to them an impedance equal to its wave impedance. Therefore, the voltage produced by the

incident wave current $I_{\max inc}$ is

$$u_{x inc} = Z_0 i_{x inc} = Z_0 I_{\max inc} \cos \omega \left(t + \frac{x}{c} \right)$$

As was shown in the preceding section, the sign of the voltage produced by the reflected wave current is always opposite to that of the current. Therefore,

$$u_{x refl} = -Z_0 i_{x refl} = -Z_0 I_{\max refl} \cos \left[\omega \left(t - \frac{x}{c} \right) \pm \varphi \right]$$

and the resultant voltage at an arbitrary point along the line is

$$u_x = u_{x inc} + u_{x refl} = Z_0 \left\{ I_{\max inc} \cos \omega \left(t + \frac{x}{c} \right) - I_{\max inc} \cos \left[\omega \left(t - \frac{x}{c} \right) \pm \varphi \right] \right\} \quad (5-35)$$

Expressions (5-33) and (5-35) enable the voltage-to-current ratio at the point x to be found

$$\frac{u_x}{i_x} = Z_0 \frac{I_{\max inc} \cos \omega \left(t + \frac{x}{c} \right) - I_{\max refl} \cos \left[\omega \left(t - \frac{x}{c} \right) \pm \varphi \right]}{I_{\max inc} \cos \omega \left(t + \frac{x}{c} \right) + I_{\max refl} \cos \left[\omega \left(t - \frac{x}{c} \right) \pm \varphi \right]} \quad (5-36)$$

As an example, we examine the case in which a quarter-wavelength line is loaded with a pure resistance. Substituting $x=0.25\lambda$ and $\varphi=0$ into the general expression (5-36), we obtain

$$\frac{u_{in}}{i_{in}} = Z_0 \frac{I_{\max inc} \cos \omega \left(t + \frac{T}{4} \right) - I_{\max refl} \cos \omega \left(t - \frac{T}{4} \right)}{I_{\max inc} \cos \omega \left(t + \frac{T}{4} \right) + I_{\max refl} \cos \omega \left(t - \frac{T}{4} \right)}$$

Since $\cos(\omega t + 90^\circ) = -\sin \omega t$ and $\cos(\omega t - 90^\circ) = \sin \omega t$, dividing the numerator and denominator by $\sin \omega t$ we find that

$$Z_{in} = Z_0 \frac{I_{\max inc} + I_{\max refl}}{I_{\max inc} - I_{\max refl}} = \frac{1-p}{1+p} \quad (5-37)$$

Substituting the reflection coefficient from formula (5-32), we have

$$Z_{in} = R_{in} = \frac{Z_0^2}{R_{load}} \quad (5-38)$$

Equation (5-36) is valid for any point along the line and for every time instant. Thus, for the time instant $t=0$ at the far end of the line, i.e. with $x=0$, where the input impedance of the line must be equal to that of its load, we can write

$$Z_{x=0} = Z_{load} = Z_0 \frac{I_{\max inc} - I_{\max refl} \cos \varphi}{I_{\max inc} + I_{\max refl} \cos \varphi}$$

Hence, the relative load impedance is

$$\frac{Z_{load}}{Z_0} = \frac{I_{\max inc} - I_{\max refl} \cos \varphi}{I_{\max inc} + I_{\max refl} \cos \varphi} = \frac{1 - p \cos \varphi}{1 + p \cos \varphi} \quad (5-39)$$

Subtracting unity from both sides of the above equation, we obtain

$$\frac{Z_{load}}{Z_0} - 1 = \frac{-2p \cos \varphi}{1 + p \cos \varphi}$$

while the addition of unity to both sides of the equation gives

$$\frac{Z_{load}}{Z_0} + 1 = \frac{2}{1 + p \cos \varphi}$$

Dividing the last two equations by each other, we find that

$$\frac{Z_{load} - Z_0}{Z_{load} + Z_0} = -p \cos \varphi \quad (5-40)$$

This formula relates the reflection coefficient with the load impedance. Since the line wave impedance is a pure resistance, the reflection coefficient can be zero only in those cases when the load impedance is also a pure resistance equal to the line wave impedance.

5-7. ACTUAL LINES WITH LOSSES

We have so far considered only ideal lines without any heat losses and leakages. With regard to their properties, actual lines are very close to ideal ones; in practice, therefore,

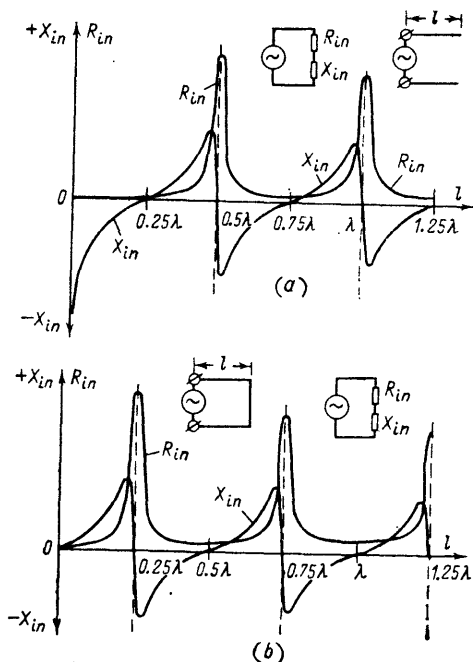


Fig. 5-19. Resistive and reactive components of the input impedance for an open and short-circuited line with losses as a function of their length

formulas for ideal lines are usually used for calculations. The shorter the line, the greater the conductor diameter and the better the wire insulation, the more precise will be the results of calculations. In some cases, however, neglecting line losses can lead to serious errors.

In the case of travelling waves, line losses cause a gradual decrease in current and voltage along the line. When standing waves are present in a line, the voltage and current in some line sections sharply rise, which results in a considerable increase of losses in the line conductors (due to conductor heating, insulation leakage, dielectric losses, etc.). As a result, the line efficiency is greatly reduced. Therefore, to trans-

smit energy it is always desirable to have a line with the travelling wave ratio as high as possible.

Because of the losses the line input impedance has always a resistive component in addition to a reactive component. For this reason, at series resonance, as in oscillatory systems with lumped elements, the line input impedance is not equal to zero, and at parallel resonance it never becomes infinitely high.

Let us consider Fig. 5-19*a* where curves for the input impedance of an open-circuited line with losses and Figs. 5-10 and 5-12 where similar curves for ideal lines are plotted, respectively, as a function of the line length. Comparison of the figures shows that the line losses result in the following effects. First, the line input impedance has a resistive component which, with the line length increasing, is found to rise at series resonance and decrease at parallel resonance. As this occurs the resonance curve of the line becomes less acute. Second, the curve for the reactive component of the input impedance of the line with losses differs only slightly from the similar curve for the ideal line, except for the curve section near resonance. On this curve section the reactance does not tend to infinity but drops abruptly to zero and then rises sharply, changing its sign (at resonance the reactance of the line is always zero). Consequently, to determine the reactive component of the line input impedance formulas (5-22) or (5-25) derived for ideal lines can be used with the exception of the above-mentioned section near the resonance.

It should be noted that at parallel resonance the line input impedance is hundreds, and sometimes thousands of times higher than at series resonance. Therefore, the load of a generator connected to the line input can sharply vary as the generator frequency changes.

In practice, feeders of various types are used (Fig. 5-20). Fig. 5-20*a* shows a two-wire balanced aerial line. The diameter of the line conductors and the spacing between them are selected depending on the power of the source to which it is connected as well as on the required wave impedance which for the lines of this type usually ranges from 300 to 800 ohms. The wave impedance Z_0 of such a line for a given spacing D between the line conductors and a given diameter d can be determined with the use of the plot in Fig. 5-21. To

protect the line conductor against corrosion, enamel-covered wires should be used. Such lines are simple and can be employed at wavelengths up to the metre range.

The operation of open (aerial) lines is highly effected by environmental conditions. For example, high humidity or icing of the line wires increase the losses in the line by several times. However, even under very unfavourable conditions

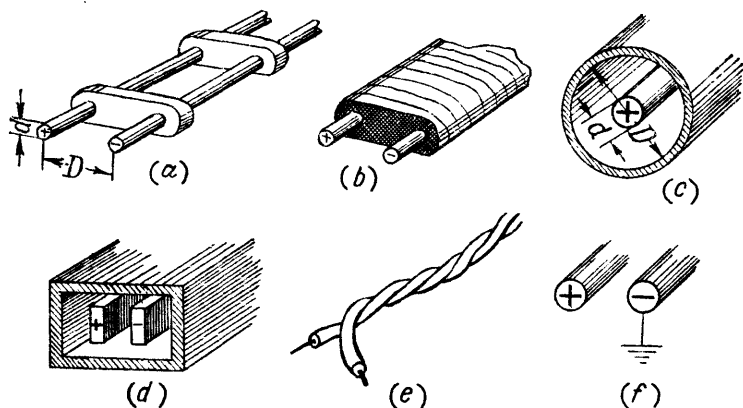


Fig. 5-20. Transmission lines of various types

the losses in the travelling wave mode are very low and can be neglected. In the standing wave mode the losses in a two-wire aerial line sometimes exceed those in the travelling wave mode by more than 5 times. Therefore, when the length of the line reaches several wavelengths, the losses have to be taken into account.

In another type of two-wire lines the wires are embedded in a dielectric (Fig. 5-20b). Such lines are used less often. They usually have a wave impedance of 50-300 ohms, which is very suitable for their matching with antennas of certain types. Another advantage of such lines is the fact that they are not influenced by environmental conditions and are easy to mount. Their main disadvantage is great attenuation caused by high losses in dielectric.

The two-wire asymmetrical line most widely used is the coaxial cable (Fig. 5-20c) in which one of the conductors is placed inside of the other, the outer conductor being made

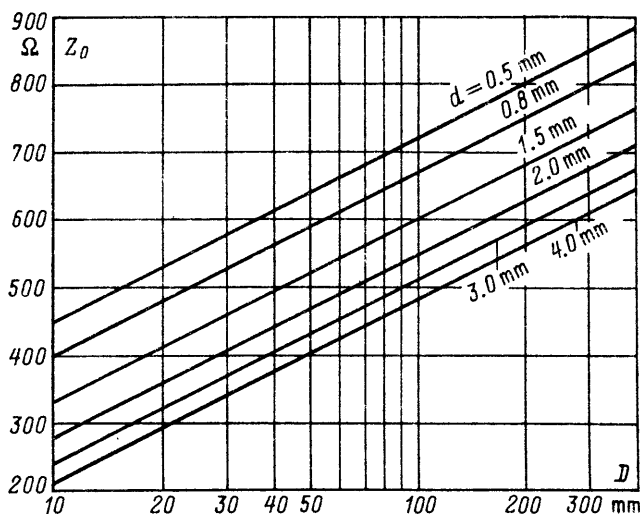


Fig. 5-21. Curves for determining the wave impedance of two-wire balanced aerial line

in the form of a flexible metal braiding. The source voltage is fed to the outer and inner conductors of the cable. The electromagnetic field produced in the coaxial cable is enclosed within the space between the conductors owing to the shielding effect of the outer conductor; this is an important advantage of the coaxial cable. The outer conductor of the cable can be grounded.

The main disadvantage of coaxial cables is high losses in the dielectric. Coaxial cables with ceramic insulators or dielectric spacers placed at some distance from one another have minimum losses.

The wave impedance of a coaxial line is

$$Z_0 = 138 \log_{10} \frac{D}{d} \quad (5-41)$$

where D is the inside diameter of the outer conductor and d is the diameter of the inner conductor.

The wave impedance of a coaxial line can be determined from the plot given in Fig. 5-22. The wave impedance of commercially available cables ranges from 40 to 150 ohms.

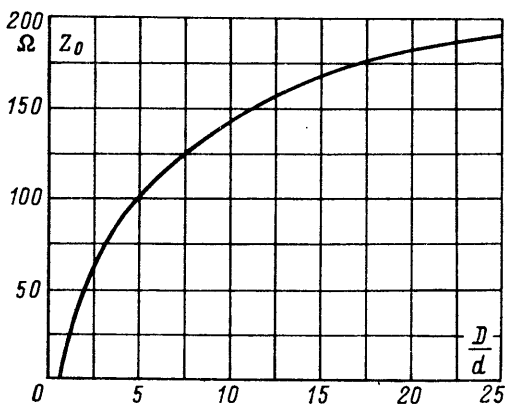


Fig. 5-22. Curves for determining the wave impedance of air dielectric coaxial lines

To transmit high power, a two-wire line is often designed in the form of two wide ribbons enclosed in a solid metal shield (Fig. 5-20d). For transmission of a relatively low po-

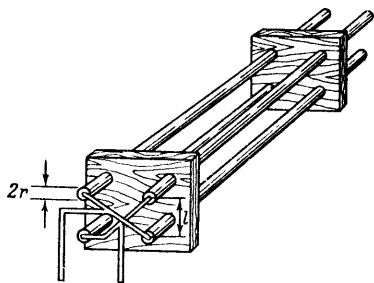


Fig. 5-23. Four-wire line

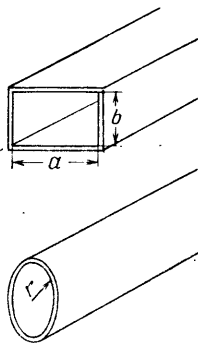


Fig. 5-24. Rectangular and circular waveguides

wer, cords of a pair of twisted flexible wires insulated with rubber are used (Fig. 5-20e). The wave impedance of such a cord usually lies in the range from 50 to 80 ohms. Such a line, however, has high dielectric losses.

In some cases it is convenient to run one of the line wires underground (Fig. 5-20f), or simply to use the ground as the

second wire. In this case the line will be asymmetrical. It should be borne in mind, however, that the simplicity of line mounting and saving of wire are achieved in this case at the cost of an increase in losses.

When a low wave impedance is required a pair of two-wire lines connected in parallel is used (Fig. 5-23). Wires disposed diagonally are connected in parallel. This enables the symmetry of the entire system to be preserved since the capacitance of the two pairs of wires relative to the ground proves to be the same. Frame insulators placed at 1.5-2 metres from one another make the entire structure rigid.

The choice of the line type is primarily determined by the line purpose, operating frequency and the amount of power being transmitted. Coaxial lines are used at wavelengths up to 10 cm. At lesser wavelengths the losses in such lines become so high that in some cases prevent their use.

At shorter wavelengths, waveguides are used. This device presents a hollow metal tube, having a rectangular or circular cross section (Fig. 5-24). Let us consider the principle of waveguide operation. Suppose that, in order to decrease losses, we build a balanced two-wire line using wide metal strips as wires (Fig. 5-25). Then we can connect the two strips to each other with a section of a quarter-wave short-circuited line without changing the operating conditions of the two-wire line since the quarter-wave line has an infinitely high input impedance. If we increase the number of such sections arranging them on both sides (top and bottom) of the line, we finally obtain a continuous metal tube. The tube's wide

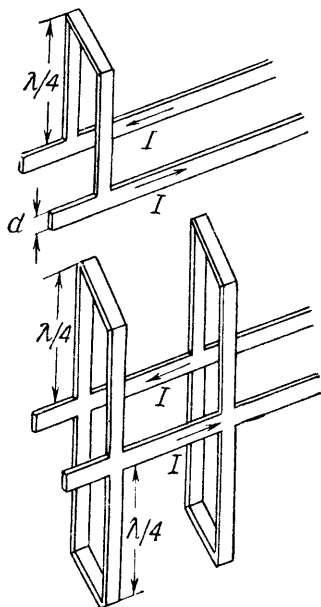


Fig. 5-25. Illustration of analogy between a rectangular waveguide and a two-wire line

walls which carry currents flowing in opposite directions serve as the main current-carrying surfaces. Since there are no supporting insulators in the tube and its conducting surface is large, the losses in waveguides are very low.

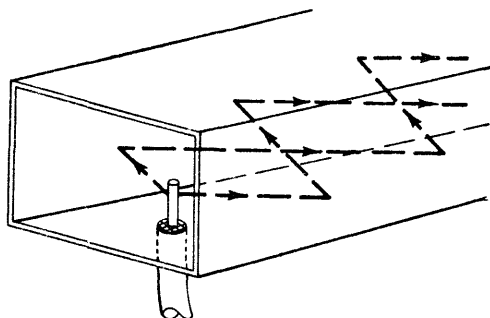


Fig. 5-26. Propagation of electromagnetic waves in a waveguide

In waveguides, electromagnetic waves are usually excited by a small radiator in the form of a short metal rod placed in the middle of the wide wall of the waveguide (Fig. 5-26). The waves radiated by the rod reach the waveguide walls setting free electrons in the bulk of the walls into oscillation. These oscillations, in their turn, produce radiation in the form of a reflected wave which travels toward the opposite wall of the waveguide, etc. Finally, after multiple reflections the electromagnetic wave reaches the far end of the waveguide.

Let us consider the phenomena which take place when an electromagnetic wave impinges on an ideally conducting surface. We assume that the incident wave is a plane wave, i.e. the surfaces passing through the points with identical phases, which are called *wavefronts*, are parallel planes (see Fig. 5-27). In this figure the traces of surfaces drawn through the points with the maximum intensity of the electrical field of the same sign are shown by dashed lines. Vector v represents the velocity of the wavefront. For air its value is numerically equal to the velocity of light, i.e. $v=c$.

Let an observer measuring the wavefront velocity be located at point A . If, at a given moment of time, the front of the wave 1 passes through point A , then the front of the

next wave 2 passes at that moment through point B ; the wavefront 2 will reach point A after time T , having passed along the plane the distance $AB = \lambda / \sin \alpha$, where α is the incidence angle of the wave. Hence, the wavefront velocity, called the *phase velocity*, for the wave moving along the plane will be

$$v_{ph} = \frac{\lambda}{T \sin \alpha} = \frac{c}{\sin \alpha} \quad (5-42)$$

Since $\sin \alpha < 1$, the phase velocity is greater than the wavefront velocity in free space. The less the angle α , the greater

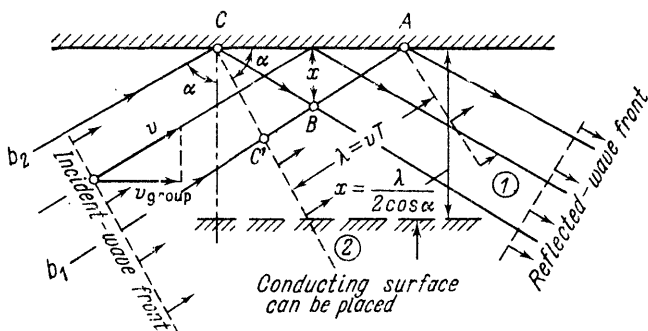


Fig. 5-27. Reflection of electromagnetic waves from an ideally conducting surface

the phase velocity. As $\alpha \rightarrow 0$, the phase velocity $v_{ph} \rightarrow \infty$, this means that the wavefront will reach points A and B almost simultaneously.

Electromagnetic waves are reflected according to the laws of geometrical optics since the physical nature of this phenomenon is the same as that of the reflection of light rays (the angle of incidence is equal to the angle of reflection). Near the surface of an ideal conductor the electric field produced in the result of addition of incident and reflected waves is normal to plane of the conductor, as the tangential component directed along the conducting plane causes a momentary redistribution of charges in the conductor, which neutralize this component with their own field. The result of this effect is that the tangential components of the electric field

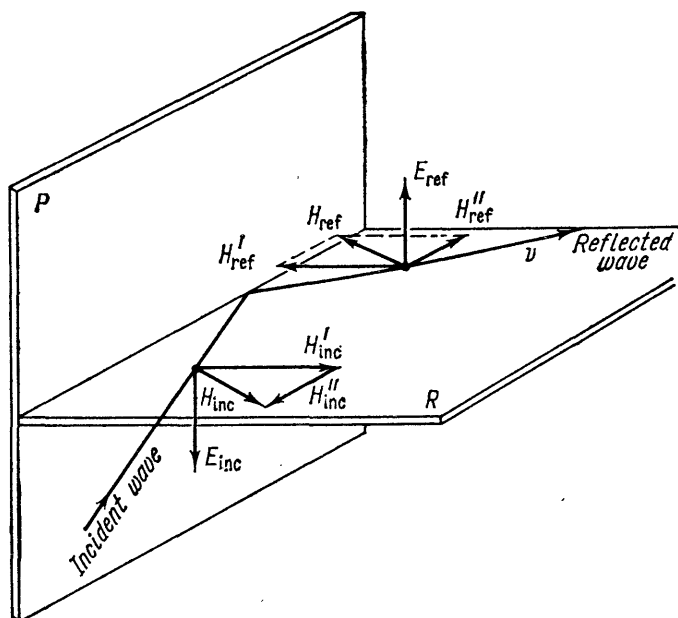


Fig. 5-28. Longitudinal and transverse components of electric and magnetic fields near ideally conducting surface

of both the incident and reflected waves are equal and opposite in direction.

If a dipole exciting the wave in the waveguide is parallel to its reflecting wall (as in the case shown in Fig. 5-26), vector E_{inc} will have the same direction. Consequently, to make the resultant tangential component of the electric field equal to zero, vector E_{refl} must have the opposite direction (Fig. 5-28). Now, using the right-hand screw rule, the direction of the vector H_{refl} of the wave magnetic field can be easily determined. In Fig. 5-28 the electric field vectors lie in the plane P on which the wave impinges, and the magnetic field vectors are in the plane R which is perpendicular to P . From this figure it can be seen that the magnetic field components perpendicular to the conducting plane of incident (H'_{inc}) and reflected (H'_{refl}) waves have opposite directions and cancel each other, while the tangential compo-

nents H''_{inc} and H''_{refl} have the same direction and enhance each other. Thus, the phase of the electric field, when the wave is reflected from the conducting plane, changes by 180° which is equivalent to the wave passing an additional distance of $\lambda/2$.

The effect of addition of the incident and reflected waves manifests itself in different ways depending on the distance from the reflecting wall, since the phase difference of the two waves is constantly changing (see Fig. 5-27). For the waves originating from the same source, for instance, the phase difference at point B will be determined by the fact that the ray $b1$ will reach it earlier than the ray $b2$ by the time $t = (CB - C'B)/c$, which will result in a phase difference $\Delta\phi_1 = 2\pi t/T$. An additional phase displacement π is caused by reflection, hence the resultant phase difference will be

$$\Delta\phi = \frac{2\pi t}{T} + \pi$$

From Fig. 5-27 it can be seen that $CB = x/\cos \alpha$, and $C'B = CB \sin (2\alpha - 90^\circ) = -x \cos 2\alpha / \cos \alpha$. Consequently,

$$t = \frac{x}{c \cos \alpha} (1 + \cos 2\alpha) = \frac{2x}{c} \cos \alpha$$

$$\Delta\phi = \pi \left(1 + \frac{4x}{cT} \cos \alpha \right) = \pi \left(1 + \frac{4x}{\lambda} \cos \alpha \right)$$

As shown above, at the reflecting surface, where $x=0$, the phase difference $\Delta\phi = \pi$. With x growing the phase difference is increased and the resultant field appears. When $x = \lambda/4 \times \cos \alpha$, the phase difference $\Delta\phi = 2\pi$, i.e. the two waves are added, and the intensity of the resultant field reaches the double intensity of the incident wave field. At the distance $x = \lambda/2 \cos \alpha$ from the reflecting surface the phase difference $\Delta\phi = 3\pi$, and the resultant field intensity equals zero. This is valid for all points spaced at the given distance from the reflecting surface. The conditions at these points are identical to those at the surface of the reflecting (conducting) plane. Therefore, the conditions of electromagnetic wave propagation will not change if a second conducting plane is placed at a distance $\lambda/2 \cos \alpha$.

The reverse is also true, *viz.*, if two parallel conducting planes exist, electromagnetic waves can propagate along them th-

rough multiple reflections provided their angle of incidence is

$$\cos \alpha = \frac{\lambda}{2a}$$

where a is the distance between the planes.

For all other angles of incidence the electric field component directed along the conducting plane is not equal to zero at its surface, therefore the waves cannot propagate.

It can be seen from the above expression that the greater the wavelength, the closer the angle of incidence to zero. If $\lambda \rightarrow 2a$, then $\alpha \rightarrow 90^\circ$, i.e. the waves propagating in the waveguide are directed almost perpendicularly to its walls, therefore they undergo a great number of reflections which results in high energy losses. When $\lambda = 2a$, the waves are normal to the side walls of the waveguide, thus they cannot propagate along the waveguide. Longer waves cannot propagate along the waveguide either. The longest wave that can yet propagate along the waveguide has the wavelength which is called *cutoff* wavelength and defined as

$$\lambda_c = 2a \quad (5-43)$$

As the wavelength approaches the cutoff value, the phase velocity of the wave in the waveguide tends to infinity; the same happens with the wavelength measured by the distance between two succeeding points where the field has the same phase. However, the rate of signal (energy) propagation along the waveguide does not increase with the phase velocity. On the contrary, due to the fact that energy propagates through multiple reflections from the side walls, the rate of energy propagation known as *group velocity* decreases. It is evident that this velocity will be determined by the rate of the wavefront propagation along the waveguide axis and will be equal to the projection upon the axis of the velocity vector of the wave impinging on the walls (see Fig. 5-27), thus we can write

$$v_{gr} = v \sin \alpha \approx c \sin \alpha \quad (5-44)$$

Comparing this expression with the formula for phase velocity, we can see that

$$v_{gr} = \frac{c^2}{v_{ph}} \quad (5-45)$$

or

$$v_{gr} v_{ph} = c^2 \quad (5-46)$$

i.e. the product of the phase velocity and the group velocity is a constant equal to the square of the velocity of light in free space, and if one of the two velocities increases, the other decreases. It should be borne in mind that the group velocity, being the projection of the velocity of the electromagnetic wave propagating in free space, can never exceed c , whereas the phase velocity can be greater than this value.

From the expressions for the group velocity and the angle of incidence the group velocity can be determined as

$$v_{gr} = c \sqrt{1 - \cos^2 \alpha} = c \sqrt{1 - \left(\frac{\lambda}{2a}\right)^2} \quad (5-47)$$

or

$$v_{gr} = c \sqrt{1 - \left(\frac{\lambda}{\lambda_c}\right)^2} \quad (5-48)$$

Whence, the phase velocity can be expressed in the following form

$$v_{ph} = \frac{c^2}{v_{gr}} = \frac{c}{\sqrt{1 - \left(\frac{\lambda}{\lambda_c}\right)^2}} \quad (5-49)$$

Multiplying both sides of this equation by the signal period T , we can obtain the following expression for the wavelength in the waveguide

$$\lambda_w = v_{ph} T = \frac{\lambda}{\sqrt{1 - \left(\frac{\lambda}{\lambda_c}\right)^2}} \quad (5-50)$$

Using (5-43) it can be shown that for $\lambda = 10$ cm the width of the waveguide must exceed 5 cm. Waveguides designed for longer waves become too bulky, heavy and expensive and, therefore, are not usually employed for wavelengths longer than 10 cm.

For practical purposes, it is important to know the structure of the electric and magnetic fields in waveguides. The wide walls where the bulk of electrical charges with opposite signs is concentrated are the main current-carrying surfaces of the

waveguide. Therefore, as in a conventional two-wire line, the lines of electric force are directed here from one wide wall to another. The direction of the lines of electric force is reversed at every half-wavelength along the waveguide

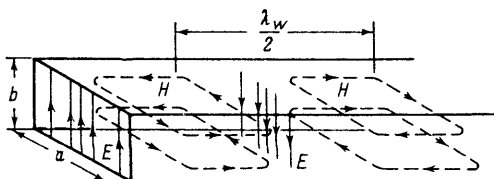


Fig. 5-29. Structure of electric and magnetic fields in a rectangular waveguide

(Fig. 5-29). The intensity of electric field in the waveguide varies over its cross section in the same way as in a short-circuited line, namely, it reaches a maximum in the centre and drops to zero at the short-circuited ends.

To explain the distribution of magnetic field in the waveguide we shall again use the analogy between a waveguide and the construction formed by addition of quarter-wave stubs to a transmission line, which was described earlier. Each of the conductors forming the stub is surrounded by its own magnetic field whose direction is determined by the right-hand screw rule. Fig. 5-30 shows these magnetic fields in a cross-section parallel to the waveguide's narrow wall. If we consider two adjacent stubs, we can see that in the space between the stubs their conductors have equal but opposite-directed fields which cancel each other. Moreover, the fields of the conductors belonging to a single stub are also mutually canceled in the space between them. Consequently, the resultant field can be represented by the lines of magnetic force surrounding the stubs.

When the number of the stubs is increased and they form continuous walls, the lines of magnetic force cannot penetrate through these walls because of their screening action; in this case they take the form of closed loops confined between the waveguide walls. The electric and magnetic lines of force reverse their direction at half-wavelength intervals along the waveguide. Thus, the magnetic field in the wave-

guide has two components, one is directed along the waveguide axis and the other is normal to it, whereas the electric field is everywhere transverse to the waveguide axis. For this reason, the waves with such a field structure are called *transverse electrical waves* and denoted by *TE*. The waves are

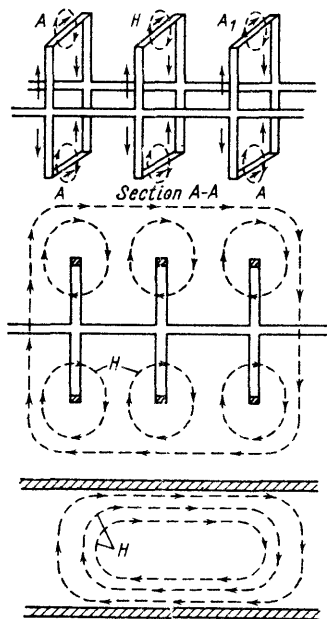


Fig. 5-30. Explanation of magnetic field structure in a rectangular waveguide

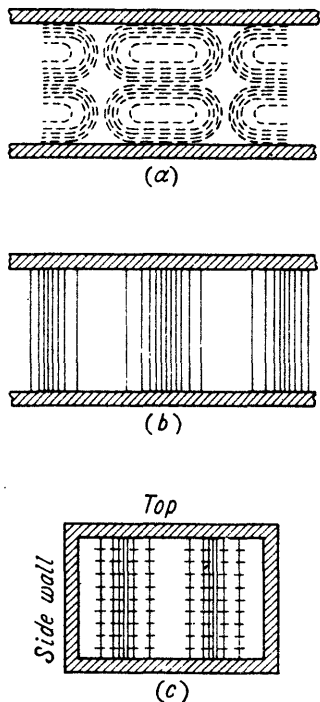


Fig. 5-31. Structure of wave $H_{0,2}$ in a rectangular waveguide

(a) magnetic field (top view); (b) electrical field (side view); (c) electrical and magnetic fields (front view)

also called after the component directed along the waveguide axis. Therefore, the above-mentioned wave is sometimes denoted by H .

As shown above, the phase difference between the incident and reflected waves is constantly changing as the distance

from the conducting surface increases. Therefore, the planes where the electric field becomes zero are spaced at equal intervals one above the other. Suppose we omit the first such plane and place the waveguide wall along the second, third, etc., then the structure of the field in the waveguide will

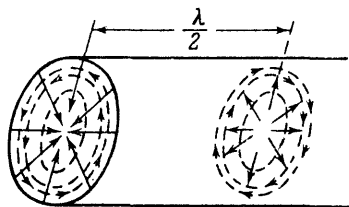


Fig. 5-32. Field structure of dominant magnetic wave in a circular waveguide (lines of electric and magnetic force are shown by solid and dashed arrows, respectively)

become more and more complicated (Fig. 5-31), and it is said that a *higher-order wave* propagates in the waveguide. The number of such waves or modes can be infinite.

In order to distinguish one type of wave or mode from another, the symbol identifying the mode is followed by two numerical subscripts. The first and second subscripts indicate the number of half-period variations in the field intensity along the larger and smaller cross-section dimensions of the

waveguide. In this notation the wave described above will be designated as $TE_{0,1}$ or $H_{0,1}$. It is also called the *dominant magnetic wave*. The wave with structure shown in Fig. 5-31 must be designated as $TE_{0,2}$ or $H_{0,2}$. Higher-order waves or modes with larger subscripts are used very rarely since the waveguides where these waves can propagate have considerably larger dimensions than those excited at the dominant frequency. Therefore, in most cases a rectangular waveguide with $TE_{0,1}$ -mode is used.

Quite rarely is use made of circular waveguides with the dominant magnetic wave whose structure (Fig. 5-32) is very similar to that of the field in a coaxial line. Circular waveguides basically serve as rotating joints and as feeders for circular horn and slot antennas. For the dominant magnetic wave in circular waveguide, the cutoff wavelength λ_c is given by

$$\lambda_c = 0.82D \quad (5-51)$$

where D is the waveguide diameter.

5-8. TRANSMISSION LINES AS REACTIVE ELEMENTS AND IMPEDANCE TRANSFORMERS

Transmission lines are used not only to transmit energy from a generator to the antenna and from the antenna to the receiver; they also find wide application as oscillatory systems, matching devices, filters and switching devices.

Quarter-wave short-circuited lines are especially suitable for use as oscillatory systems, since parallel resonance in such lines can be achieved with the line of minimum size. For example, they perform this function in ultrashort wave generators. Owing to small losses, such lines have a higher quality factor than lumped element circuits. In the decimeter wave range, two-wire lines are replaced by coaxial ones, as in the former the spacing between conductors becomes comparable with the wavelength, which results in high radiation losses.

Let us examine the use of transmission lines as matching devices. In order to transmit energy through the line without reflection, the line load must be a pure resistance equal to the wave impedance of the line. In practice, the input impedance of an antenna often differs from the wave impedance of the line. Therefore, the line cannot be matched directly to the antenna and special matching devices have to be used.

The simplest matching device is the so-called *tuned transformer* (Fig. 5-33a). A short-circuited or open-circuited line (a stub) whose length is close to a quarter wavelength is connected to the line load, and the main supply line is connected to the stub.

First, the load—matching line system must be tuned. If the load is a pure resistance, the length of the matching line is first set at 0.25λ . If the load has a reactive component, the length of the matching line is chosen slightly more or less than 0.25λ . In this case the stub becomes equivalent to a reactive load whose reactance is selected to be equal in magnitude and opposite in character to the reactive component of the load impedance. If this condition is met, the line together with its load becomes equivalent to a parallel tuned circuit (Fig. 5-33b), and the load (which is now the combination of the original load and the matching line) is a resonant system whose input impedance is a pure resistance. Then, the

main supply line must be connected to the load in such a way that the input impedance of the loaded matching line be equal to the wave impedance of the supply line.

In the matching line standing current and voltage waves are set up with a current loop and a voltage node near the line

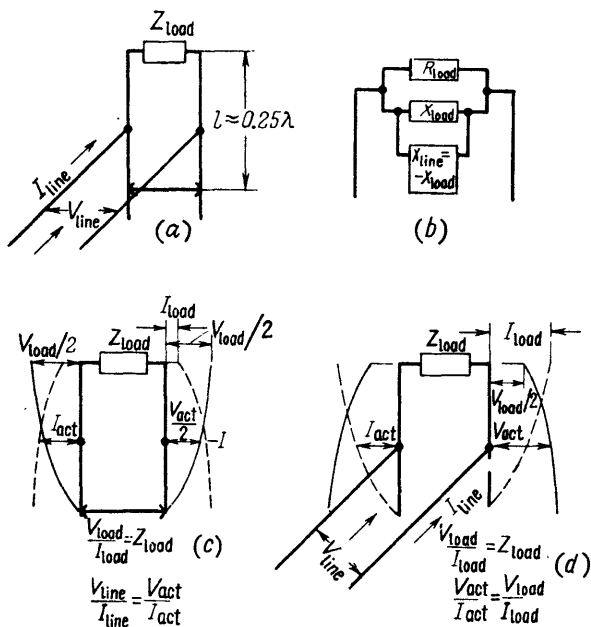


Fig. 5-33. Tuned impedance transformers

short-circuited end. As the line length is close to 0.25λ , a current wave node and a voltage loop appear in the line load position (Fig. 5-33c). At this point the ratio of voltage to current by Ohm's law must be equal to the load impedance. When moving along the line from its load to the short-circuited end, this ratio decreases and turns to zero at the opposite end of the line. If the load impedance is greater than the wave impedance of the supply line, then there will be a point along the matching line, where the ratio of voltage to

current (input impedance) will be precisely equal to the wave impedance of the supply line. If we connect the line to the load at this point, they prove to be matched.

The matching line is essentially lossless. Therefore, all the energy transmitted without reflection through the line is consumed in its load. When the load impedance is less than the line wave impedance a quarter-wave open-circuited stub must be used (Fig. 5-33d). When moving along the stub toward its open-circuited end, the stub impedance grows gradually tending to infinity. Thus, such a stub will transform the load impedance, making it higher.

Such impedance-transforming lines are often called *impedance* or *matching transformers*. Tuned transformers perform well provided their transformation ratio exceeds four. With lesser transformation ratios the feeder input must be placed near the loaded end of the transforming line. In this case the feeder and the load strongly effect each other, and it becomes difficult to find the correct position for the connection of the feeder to the line. In such cases, a reactive stub should be used as a matching element.

When a travelling wave combined with a standing one is set up in a loaded line, the input impedance of the line changes from point to point along its length. At a certain point the resistive component of this impedance is found to be equal to the line wave impedance. Therefore, to obtain the impedance match, it is necessary to compensate for the reactive component of the line input impedance at this point. For this purpose a short- or open-circuited stub (Fig. 5-34) whose reactance is equal in magnitude and opposite in sign to the reactive component of the line input impedance is connected to the line at this point. When it is done, only a travelling wave will propagate along the line, and all the energy supplied by the generator will be transmitted to the load with no reflection.

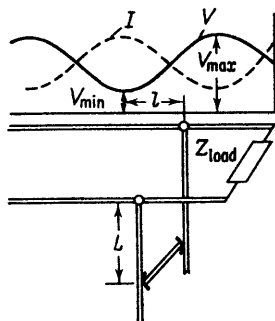


Fig. 5-34. Matching a line to a load with the aid of a reactive stub

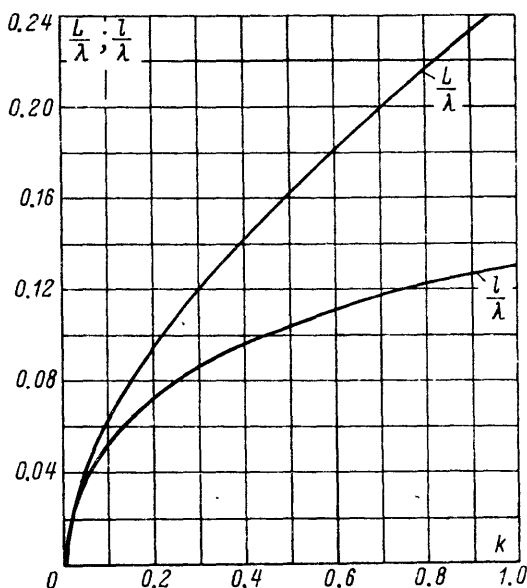


Fig. 5-35. Curves for determining the length of a stub and its position along the line

The correct position of the stub along the line and the stub length can be determined with the use of Fig. 5-35. First, the line's travelling wave ratio is measured (see below). Then, from the plot shown in the figure, the ratio l/λ is determined, which gives the distance between a voltage node and the point where the stub should be connected to the line. From the second curve given in Fig. 5-35, the stub electrical length L/λ is next determined. Although the stub can be placed near any voltage node, to decrease losses it is advisable to place it at the node nearest to the line far end. The stub length must be minimal. In order that the stub length be less than $\lambda/4$ the short- and open-circuited stubs should be used respectively for the inductive and capacitive input impedances of the line.

The character of the line input impedance at a given point can be determined from the distribution of voltage along the line. If the stub connection point is on the descending

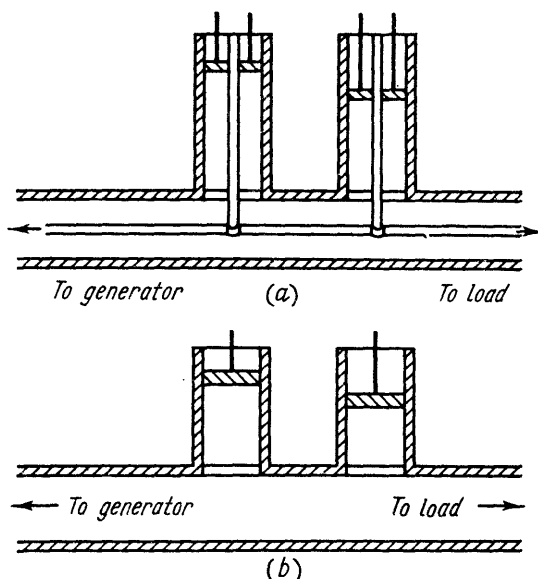


Fig. 5-36. Application of two stubs as impedance matching transformers
(a) in a coaxial line; (b) in a waveguide

portion of the voltage curve (as measured from the far end of the line), the reactive component of the input impedance of the line section which follows this point will be of a capacitive character. If, however, the stub connection point appears on the ascending portion of the voltage curve, the line reactance at this point will be of an inductive character.

Since the construction of a movable stub for coaxial lines presents some problems, two reactive stubs are used with such lines (Fig. 5-36a). By adjusting the length of the first stub, i.e. the one closest to the line load, the resistive component of the input impedance of the right-hand line section is made equal to the wave impedance of the supply line. This done, a certain reactive component still remains at the input. To compensate for this component a second stub is usually placed at a short distance from the first, with the length of the second stub being also adjustable. Thus a match can be achieved with the aid of the system consisting of two

stationary stubs. Similar devices are also employed in waveguides (Fig. 5-36b).

To match a transmission line to a pure resistive load (a small reactance can be compensated for in the same way as

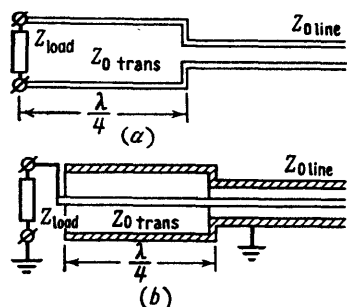


Fig. 5-37. Quarter-wave transformers

(a) in a balanced two-wire line; (b) in a coaxial cable

with the transformer of the first type), the so-called *quarter-wave transformers* are used (Fig. 5-37). In Section 5-6 it was shown that a quarter-wave line loaded with a resistance R_{load} has an input resistance $R_{in} = Z_{0\ Tr}^2 / R_{load}$, where $Z_{0\ Tr}$ is the wave impedance of the transformer. By selecting the wave impedance of the transforming line, its input impedance can be made equal to the wave impedance of the supply line. The wave impedance of the trans-

forming line is then given by

$$Z_{0\ Tr} = \sqrt{R_{load} Z_{0\ line}} \quad (5-52)$$

where $Z_{0\ line}$ is the wave impedance of the line being matched.

When the load resistance cannot be determined directly, the load is connected to the supply line and the line's traveling wave ratio $k = R_{load} / Z_{0\ line}$ is measured. Whence, the load resistance is

$$R_{load} = k Z_{0\ line}$$

Substituting this expression into (5-52), we obtain the formula for the transformer wave impedance

$$Z_{0\ Tr} = Z_{0\ line} \sqrt{k} \quad (5-53)$$

To tune the transformer to different wavelengths, it is often built in the form of telescopic tubes.

Quarter-wave transformers also find application in coaxial lines, though in this case they become more complicated (Fig. 5-37b). Moreover, in practice it is very difficult to achieve an ideal match because of reflections from a discontinuity presented by the line-to-transformer joint. In this case a match

is accomplished with the aid of movable field meters loosely coupled with the line. A radio-frequency instrument mounted on a light-weight cross-member sliding along the line can serve as such a meter (Fig. 5-38). The instrument is usually capacitive coupled to the line. When a very high sensitivity is required, a dc microammeter connected into the circuit of a semiconductor detector is used. The readings of such an instrument are proportional to the voltage across the points along the line at which the instrument is placed. Moving the instrument along the line and noting its minimum and maximum readings, one can determine the line's travelling wave ratio and the distribution of standing waves along the line.

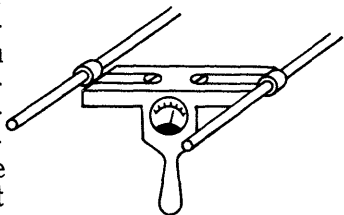


Fig. 5-38. Instrument for measuring electrical field distribution in a balanced two-wire line

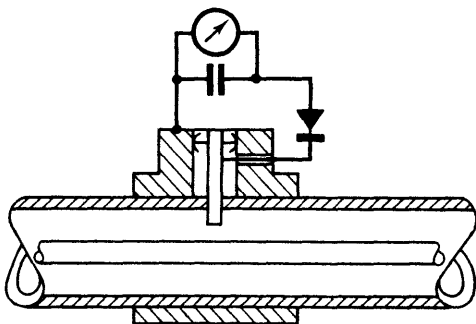


Fig. 5-39. Instrument for measuring electrical field distribution in a coaxial line

For coaxial lines and waveguides, special measuring instruments are used (Fig. 5-39). To make measurements, the line or waveguide being tested is cut and the instrument is connected between the two sections. Such instruments include sections of coaxial lines made of rigid tubes or waveguides along which a narrow slot is cut. A small probe extends into the guide through the slot. This probe serves as a small

receiving antenna; the signal from the probe is applied through the detector to the meter. By moving the probe along the slot, the field distribution along the line or waveguide can then be obtained. The measurement results will be accurate only if the measuring instrument is matched to the line being tested.

In short-wave radio stations, transmission lines are also employed as filters. In general a filtering device must transmit the waves of length λ_1 and attenuate the waves of length λ_2 . This can be accomplished by connecting two stubs to the

main line as shown in Fig. 5-40. The open-circuited stub $0.25 \lambda_2$ long acts as a short circuit for the wave λ_2 which propagates along the line from left to right. Therefore, this wave is completely reflected from the points AA. If λ_1 is greater than λ_2 , the first stub presents a capacitive reactance to the wave λ_1 . By connecting a short-circuited inductive stub in parallel with the first one and matching their impedances, we obtain a parallel resonance circuit which presents an infinitely high resistance to the wave λ_1 . Therefore, the shunting action of the second stub with respect to the wave λ_1 can be neglected and one can say that all input energy is transmitted through the filter to the load.

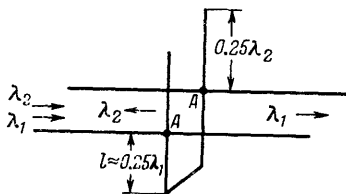


Fig. 5-40. Application of two stubs as a filter

6-1. RADIATING SYSTEMS

Static electric charges produce a static electric field in the surrounding space, whereas a direct current sets up a dc magnetic field. Alternating current flowing in a pair of conductors connected to the terminals of an alternating emf source generates an alternating electric field propagating in space.

The intensity of radiation field may be different depending upon the shape of conductors. When mounting the feeder lines through which energy is transmitted from a generator to a load, designers try to reduce radiation in every possible way. For this purpose, the line wires are arranged in parallel and close to each other. In this case the fields produced by two equal but opposite-directed currents (flowing in the line wires), i.e. the currents which are opposite in phase, cancel each other and no energy is radiated into the surrounding space. Radiation is eliminated most completely when a coaxial cable is used as the electric field is contained in the space between the inner conductor and the inner surface of the outer conductor.

A quite opposite task is assigned, i.e. the production of the maximum possible radiation, when designing antennas. For this purpose, the transmission lines previously discussed can be used after eliminating the main reason which deprive such feeders of their radiating properties. This can be achieved by several ways. For instance, line wires can be placed at some angle so that their fields cannot cancel each other. This principle underlies the operation of *V-shaped and rhombic antennas* (Fig. 6-1a and b) whose radiating wires are

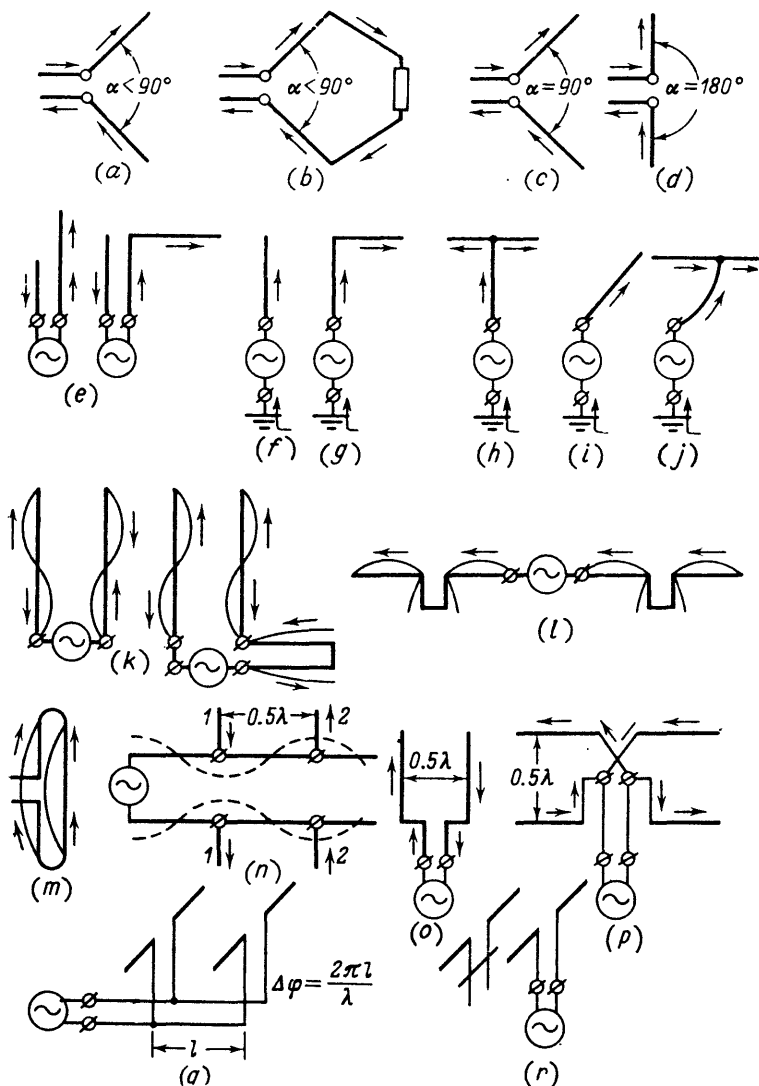


Fig. 6-1. Principal types of radiating devices

arranged at a sharp angle to one another, the Pistolcors' *V-antenna* (Fig. 6-1c) where wires are at a right angle (90°) to one another and a *double-dipole antenna* formed by two wires spaced through 180° (Fig. 6-1d). In Fig. 6-1 the direction of currents flowing through wires is indicated by arrows.

One of the feeder wires can be made shorter or eliminated from the system altogether. In this case, the fields produced by the currents in the two wires will not cancel each other completely. One of the wires was shortened for the first time in the so-called "*top-fire*" antennas suggested by M. A. Bonch-Bruyevich (Fig. 6-1e). The so-called *asymmetrical dipole* (Fig. 6-1f) is obtained by eliminating the second wire altogether. All antennas using this principle fall into the category of asymmetrical antennas. Among them are inverted *L-antennas* (Fig. 6-1g), *T-shaped antennas* (Fig. 6-1h), *tilted antennas* (Fig. 6-1i), a *tapped horizontal antenna* (Fig. 6-1j), etc. It is common practice to ground the second terminal of the generator feeding such antennas (the ground serves as the second wire).

A feeder radiates into space if adjacent sections of its two wires carry the currents which are in phase, thus their fields reinforce each other. For this, it is sufficient to bend a half-wavelength section of one of the wires into a nonradiating stub (Fig. 6-1k). The so-called *cophased antennas* based on this principle (Fig. 6-1l) as well as the *folded dipole antennas* suggested by A. A. Pistolcors (Fig. 6-1m) have extensive applications in radio systems.

The feeder radiates if the distance between its wires is comparable with the wavelength. In this case the waves radiated by the two wires in some directions acquire an appreciable path-length difference and do not cancel each other. Furthermore, the distance between the wires may be selected so that in some directions summation of the waves emanating from the two wires would take place. This method is widely used in a variety of the so-called *antiphased antennas*. The three examples given in Fig. 6-1 illustrate the operation of such antennas. In the antenna shown in Fig. 6-1n the antiphase currents in wires 1 and 2 are ensured by connecting them to the feeder at a distance of $\lambda/2$ from each other. The antenna in Fig. 6-1o looks as if the end sections of the feeder wires are spaced more apart. In the antenna shown in Fig. 6-1p

the antiphase currents are provided by transposition of the feeder wires.

Antenna radiation can be increased by changing the phase of currents in the parallel wires and the distance between them. In a system consisting of several double-dipoles a desired phase shift between the currents in the dipoles can be, for instance, obtained by selecting the distance between the points where these dipoles are connected to the feeder as in Fig. 6-1*q* or by tuning one dipole with the aid of a stub in which current is induced by another dipole as shown in Fig. 6-1*r*.

Antennas which simultaneously utilize several of the principles indicated above are also widely encountered in radio equipment.

6-2. DOUBLE-DIPOLE ANTENNAS

A double dipole may be represented as an open-circuited transmission line whose wires are spaced through 180°

(Fig. 6-2). Owing to this construction, the current directions at symmetrical points of both halves of the dipole prove to be identical. To obtain the maximum current in the dipole wires and, consequently, the maximum radiation the dipole is tuned to resonance, i.e. the frequency of the current in the dipole is set equal to that of the dipole free oscillations.

Let us consider the process of free electric oscillations in a double dipole. For this, we connect its both halves to the terminals of a dc voltage source (Fig. 6-3*a*). As soon as the distributed capacitances of the dipole wires are charged and a potential difference is set up between the dipole halves, the supply source is disconnected and both dipole halves are connected with a short jumper (Fig. 6-3*b*). As this takes place, the distributed capacitances start to discharge through the

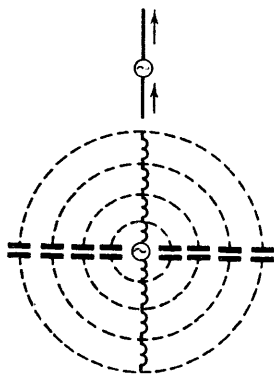


Fig. 6-2. Double dipole and its equivalent circuit

potential difference is set up between the dipole halves, the supply source is disconnected and both dipole halves are connected with a short jumper (Fig. 6-3*b*). As this takes place, the distributed capacitances start to discharge through the

jumper. It is evident that the maximum electric charge will flow through the sections of the dipole wires located at the dipole centre and, therefore, the discharge current attains its maximum value in these sections, whereas at the wire ends it drops down to zero. The current increases gradually due to

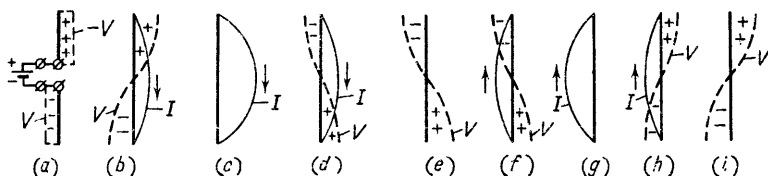


Fig. 6-3. Free oscillations in double dipole

the action of an emf self-induced in the wire distributed inductances. A potential difference between the points equally spaced from the centre of the dipole is the higher, the greater their distance from the dipole centre since the greater part of the wire distributed inductance produces this difference (Fig. 6-3b). The potential sign with respect to the dipole centre tap is different on either side since in one half of the dipole current flows towards the centre tap, whereas in the other half, away from it.

As the distributed capacitance discharges, the current in the wire rises and attains its maximum when the capacitance is fully discharged. As this takes place, all the electric field energy stored by the capacitance is converted into the magnetic field energy stored in the distributed inductance (Fig. 6-3c). If at the beginning the dipole inductance has tended to prevent a current increase, now it tends to prevent its decrease. Therefore, the current decreases slowly maintaining its previous direction (Fig. 6-3d). As this takes place, the distributed capacitance recharges and proves to be fully recharged (Fig. 6-3e) when current drops down to zero, after this the process occurs in the reverse direction (Fig. 6-3f, g, h, i).

Thus, free electric oscillations occur in the double dipole, and current and voltage standing waves are set up along it, with a half-period variation occupying the entire length of the dipole. Consequently, the wavelength λ_0 of free oscil-

lations in a double dipole is twice as great as its length l , i.e.

$$\lambda_0 = 2l \quad (6-1)$$

Therefore, such a dipole is also known as a *half-wave dipole* to emphasize the fact that it is twice as short as the wavelength of its free oscillations.

When a transmission line is converted into a radiating system, i.e. an antenna, the net losses increase since in addition to heat losses P_h radiation losses P_Σ arise. Thus, the power consumed by the antenna is

$$P_A = P_h + P_\Sigma \quad (6-2)$$

This increase in losses can be considered as an increase in the antenna resistance, i.e. the latter quantity may be expressed as

$$R_A = R_h + R_\Sigma \quad (6-3)$$

where R_h is the heat loss resistance, and R_Σ is the so-called *radiation resistance*.

In the systems with distributed parameters where current and voltage vary from point to point it is advisable to determine the system resistance from energy considerations using the Joule—Lenz law which relates power to current and resistance, i.e. $P = I_{\max}^2 R / 2$. Let us assume that the antenna resistance is equal to a lumped resistance in which the same power as in the antenna is dissipated when the current equal to the maximum antenna current flows through it. Then

$$R_A = \frac{2P_A}{I_{\max A}^2} = \frac{2P_h}{I_{\max A}^2} + \frac{2P_\Sigma}{I_{\max A}^2} \quad (6-4)$$

where $I_{\max A}$ is the antenna current at a standing-wave loop.

From expression (6-3), we can write

$$R_h = \frac{2P_h}{I_{\max A}^2} \quad \text{and} \quad R_\Sigma = \frac{2P_\Sigma}{I_{\max A}^2} \quad (6-5)$$

Since any antenna serves as an energy converter, its efficiency η_A equal to the ratio of the radiation power to the input power is the most important characteristic of the antenna:

$$\eta_A = \frac{P_\Sigma}{P_\Sigma + P_h} \quad (6-6)$$

Using expressions (6-6) and (6-5), we obtain

$$\eta_A = \frac{R_\Sigma}{R_\Sigma + R_h} = \frac{1}{1 + \frac{R_h}{R_\Sigma}} \quad (6-7)$$

From this equation it follows that the greater the radiation resistance as compared with the heat loss resistance, the greater is the antenna efficiency. The magnitude of the radiation resistance, however, is also of great importance. As in the

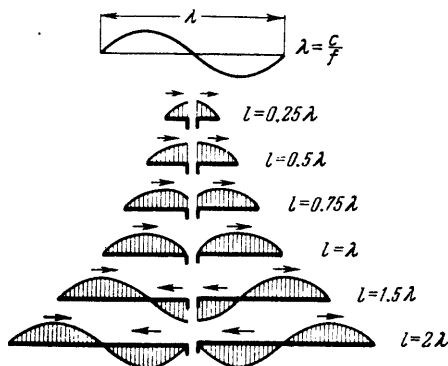


Fig. 6-4. Current distribution in double dipoles of different length

case of any oscillatory system, current and voltage at resonance as well as antenna resonance properties depend essentially upon its resistance. In fact, for a given input power the greater the antenna resistance, the less the antenna current at resonance and the less the voltage across the antenna. The last fact is advantageous since requirements for antenna insulation can be loosened.

Antenna radiating properties and, consequently, its radiation resistance depend upon current distribution in the radiating wire. In case of forced oscillations in the antenna their frequency is determined by the frequency of the voltage applied to the antenna. Current and voltage distribution along the antenna wire can vary drastically. This follows from the fact that the current and voltage waves excited in the antenna are reflected from the far end of the antenna wire and form standing waves as in the case of an open-circuited

transmission line. To find the current and voltage distribution under conditions of forced oscillations, it is necessary to determine the length of the wave being excited, using formula (5-1), for a given oscillator frequency and velocity of propagation of electromagnetic waves along the wire. Then on a scaled drawing of the antenna the standing wave distribution should be plotted using the open-circuited end of the antenna wire where a current node and voltage loop are located, as a reference point (Fig. 6-4).

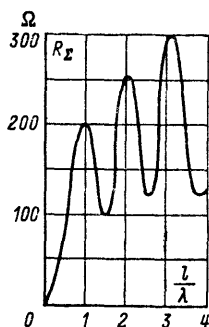


Fig. 6-5. Relationship between radiation resistance and electrical length of double dipole

The antenna radiation resistance gradually rises (Fig. 6-5) as the dipole length varies from zero to $l=\lambda$. If the dipole length $l>\lambda$, i.e. a half-period variation and a part of the next one are along each half of the dipole, currents in some sections of the antenna wire have an opposite direction with respect to those flowing through the extreme sections. Radiation from these sections weakens partially that from the extreme sections and, therefore, the radiation resistance is somewhat reduced. The radiation resistance, however, rises again with the increasing dipole length and attains a new maximum value when the dipole length $l=2\lambda$, i.e. when the radiations from the sections carrying antiphase currents would seem to cancel each other.

Let us explain this phenomenon. As seen from Fig. 6-6, two waves emanating from the dipole sections carrying opposite-directed currents reach sufficiently remote points located on the perpendicular to the dipole axis in antiphase (Fig. 6-6a) and cancel each other. These waves, however, come to the points located on the straight lines running at lesser angles to the dipole axis (Fig. 6-6b) with an appreciable path difference $\Delta=d_2-d_1=d_4-d_3$. Therefore, the difference in their phases may approach zero or 360° and antenna radiation may increase in these directions. This explains why the radiation resistance becomes higher when $l=2\lambda$. Similar changes occur as the antenna length is increased further: the radiation resistance increases and decreases alternately with

Let us explain this phenomenon. As seen from Fig. 6-6, two waves emanating from the dipole sections carrying opposite-directed currents reach sufficiently remote points located on the perpendicular to the dipole axis in antiphase (Fig. 6-6a) and cancel each other. These waves, however, come to the points located on the straight lines running at lesser angles to the dipole axis (Fig. 6-6b) with an appreciable path difference $\Delta=d_2-d_1=d_4-d_3$. Therefore, the difference in their phases may approach zero or 360° and antenna radiation may increase in these directions. This explains why the radiation resistance becomes higher when $l=2\lambda$. Similar changes occur as the antenna length is increased further: the radiation resistance increases and decreases alternately with

its value gradually rising at succeeding maximums and minimums.

It is evident from the above discussion that the radiation resistance does not depend upon the absolute dimensions of a dipole but is determined by their ratio to the wavelength. For instance, all half-wave dipoles have the radiation resistance $R_{\Sigma}=73.1$ ohms irrespective of the fact whether their

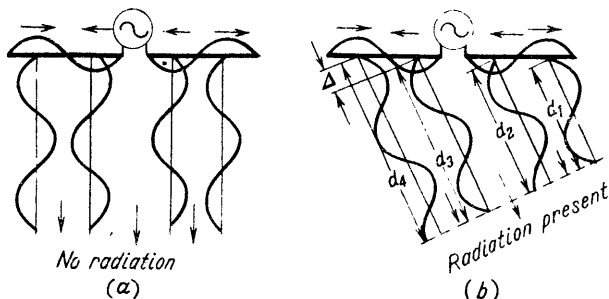


Fig. 6-6. Addition of the waves radiated from separate sections of double dipole in different directions (d — path length, Δ — path difference)

length is equal to hundreds of metres or several centimetres. All the dipoles with their length equal exactly to the wavelength have the radiation resistance $R_{\Sigma}=200$ ohms, etc.

In addition to a standing wave, a travelling wave is also produced in any antenna due to radiation and heat losses occurring in it. In all antennas (except for travelling-wave antennas), however, the amplitude of the travelling wave is significantly less than that of the standing wave and it may be, therefore, neglected when calculating the reactive component of the antenna input resistance.

To a first approximation, the distribution of current and voltage standing waves in an antenna is the same as in an open-circuited line, whence

$$I_l = I_{\max A} \sin \frac{2\pi l}{\lambda} \quad \text{and} \quad V_l = V_{\max A} \cos \frac{2\pi l}{\lambda}$$

where $I_{\max A}$ and $V_{\max A}$ are current and voltage at a loop of standing waves in the antenna wire, and l is the distance from its open end. Thus, the reactive component of the an-

tenna input impedance is defined as

$$X_{in A} = -\frac{V_{\max A}}{I_{\max A}} \cot \frac{2\pi l}{\lambda} = -Z_{0A} \cot \frac{2\pi l}{\lambda} \quad (6-8)$$

By analogy with transmission lines, the ratio of the voltage to current in an antenna we shall call the wave impedance of the antenna wire.

When designing an antenna, it is very important to know the value of its input impedance at resonance. In this case, the antenna input impedance will be a pure resistance. The power consumed by an antenna is

$$P_A = \frac{1}{2} I_{in A}^2 R_{in A} \quad (6-9)$$

whence

$$P_{in A} = \frac{2P_A}{I_{in A}^2} \quad (6-10)$$

At resonance a current loop will be at the antenna feed point provided the dipole length is equal to 0.5λ , 1.5λ , 2.5λ , etc. Consequently $I_{in A} = I_{\max A}$ and in accordance with expression (6-3) we have

$$R_{in A} = R_{\Sigma} + R_h \quad (6-11)$$

This resistance is relatively low (from ohms to tens of ohms).

The nature of antenna input impedance variation, as the ratio l/λ varies in the vicinity of 0.5 , 1.5 , 2.5 , etc. (Fig. 6-7), corresponds to impedance variation near resonance in a series oscillatory circuit. Therefore, it is often said that in the antenna series resonance takes place at a given ratio l/λ . For a dipole length equal to λ , 2λ , 3λ , etc., the antenna is fed at a voltage loop, thus the power consumed by the antenna will be

$$P_A = \frac{1}{2} I_{\max A}^2 R_A = \frac{1}{2} \frac{V_{in A}^2}{R_{in A}} = \frac{1}{2} \frac{V_{\max A}^2}{R_{in A}} \quad (6-12)$$

whence

$$R_{in A} = \left(\frac{V_{\max A}}{I_{\max A}} \right)^2 \frac{1}{R_A} \quad (6-13)$$

As the ratio of voltage to current in the antenna is equal to its wave impedance, we have

$$R_{in A} = \frac{Z_{0A}^2}{R_A} \quad (6-14)$$

In conventional wire antennas $Z_{0A} \approx 1\,000$ ohms, whereas R_A ranges from ohms to tens of ohms and $R_{in A}$ amounts to thousands and even dozens of thousands of ohms. As can be seen from Fig. 6-7, when the ratio l/λ varies in the vicinity of integers the antenna input impedance varies similarly to the impedance of a parallel oscillatory circuit near resonance. Therefore, it is often said that parallel resonance occurs at such values of the ratio l/λ .

The operating conditions of an antenna when it is fed at a current or voltage loop differ drastically. In the first case, the antenna constitutes a resistive load with a low resistance, consuming a comparatively large current with a low input voltage. In the second case, the antenna offers a very high input resistance, consuming a very small current with a high input voltage. Therefore, it is often said that in the first case the antenna is current fed and in the second case, voltage fed.

Comparison of the curves given in Fig. 6-7 with similar curves for an actual transmission line (refer to Fig. 5-19) shows that the phenomena occurring in both systems are quite similar. It should be noted, however, that there are a number of distinguishing features: (1) the resonance curves of antennas are much less sharp than those of transmission lines; (2) the resistive component of antenna input impedance is much greater and much lesser at series and parallel resonances, respectively; (3) the maxima of the antenna resonance curve decrease sharply and become less distinct with the increasing antenna length; (4) the accurate resonance length of an antenna is by several per cent less than an integral number of the half-wave. These features are explained by the

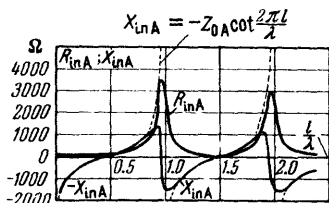


Fig. 6-7. Relationship between input resistance of double dipole and its electrical length

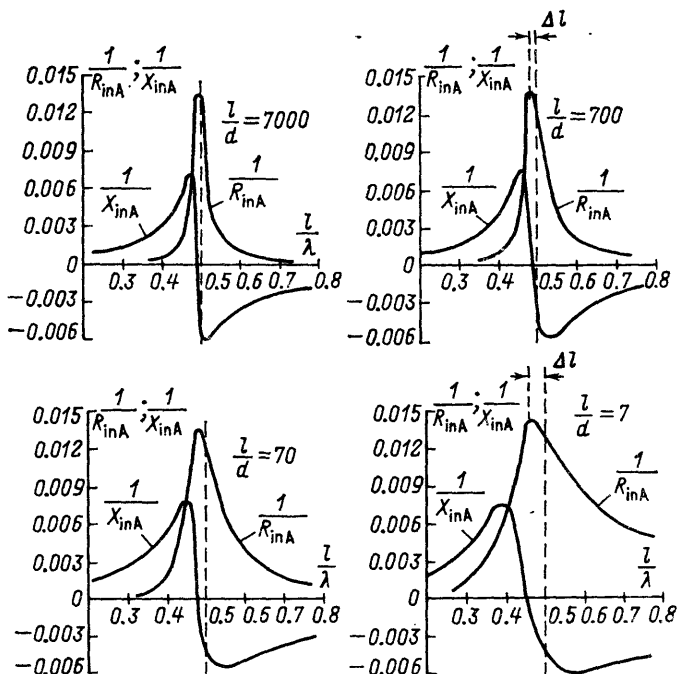


Fig. 6-8. Resonance curves for half-wave dipoles having different diameter

fact that losses due to radiation in an antenna are much higher than those in a conventional non-radiating line.

As seen from resonance curves for half-wave dipoles with different wire diameters shown in Fig. 6-8, the accurate resonance length of a dipole proves to be somewhat less than 0.5λ . This is explained by some change in the distribution of standing waves in the dipole as compared with the transmission line which has no radiation losses. As a result, an inductive component of input impedance appears in the antenna whose length is equal to an integral number of the half-wave. Therefore, the dipole should be somewhat shortened to compensate for this inductive component.

Antenna resonance properties depend essentially on its wire diameter. The antenna wave impedance is reduced with the increasing diameter. It means that its input impedance also

changes. Therefore, the shortening length required to compensate for the reactive component of the input impedance must be also changed. This is confirmed by the antenna resonance curves plotted in Fig. 6-8 for different wire diameters. These curves indicate that the length of dipole shortening increases with the increasing dipole wire diameter.

The length of dipole shortening is given by the formula

$$\Delta l \approx 0.259 \frac{\lambda}{\log_{10} \frac{\lambda}{2d}} \quad (6-15)$$

where Δl is the shortening length for a half-wave dipole, and d is the dipole wire diameter.

Symmetry of the antenna system should be maintained when shortening the half-wave dipole, i.e. both halves of the dipole should be shortened by $\Delta l/2$. Conventional single-wire antennas are usually shortened by about 5 per cent of their length. Therefore, the resonance length of a half-wave dipole can be determined from the formula

$$l_{res} \approx 0.95 \frac{\lambda}{2} = \frac{143}{f} \quad (6-16)$$

where l_{res} is the resonance length in metres, and f is the operating frequency in megacycles.

The accurate resonance length of dipoles also depends upon the number and position of insulators, antenna-wire insulation and the presence of close objects (guy ropes, posts, reinforced concrete buildings, etc.) introducing additional losses into the antenna. Therefore, the final tuning of an antenna is generally accomplished on site.

The resonance curves shown in Fig. 6-8 also indicate that the greater the wire diameter the greater the antenna bandwidth. This fact is frequently used in practice. For instance, to receive television programs which have a very wide band of frequencies, the antennas fabricated of large-diameter wires are used.

If a transmitter is designed to operate in a wide band of frequencies, we have to deal with an untuned antenna. To obtain resonance, the reactive component of antenna input impedance is balanced out with the aid of tuning controls. As this takes place, the resistive component can be determi-

ned from the power consumed by the antenna,

$$P_A = \frac{1}{2} I_{in A}^2 R_{in}$$

where the current at the antenna input is determined from the current standing-wave equation:

$$I_{in A} = I_{\max A} \sin \frac{2\pi l}{\lambda}$$

Consequently,

$$R_{in} = \frac{2P_A}{I_{\max A}^2 \sin^2 \frac{2\pi l}{\lambda}}$$

Comparing this expression with formula for the antenna resistance (6-4), we find that

$$R_{in} = \frac{R_A}{\sin^2 \frac{2\pi l}{\lambda}} \quad (6-17)$$

Formula (6-17) holds for the antennas of any length except for the cases of parallel resonance when the antenna feed point appears at a current node. In this case formula (6-14) should be used.

Antennas of different shape and size produce radiation of different intensity at the same value of loop current. To simplify the comparison of antennas of different designs one more parameter, the effective length of an antenna, is introduced. As said earlier, power radiated from an antenna wire section is proportional to the square of the current in this section. Consequently, the radiation intensity is proportional to the current. Intensity of the antenna resultant field can be determined graphically by plotting to a certain scale a current distribution along the antenna wire (Fig. 6-9a). On dividing this plot into small sections (within which the current may be assumed constant), their length is multiplied by the value of the current in each section and after summing up the results, the area $S_{1\text{curr}}$ under the current curve is determined. In short, this area is referred to as an antenna current area.

The same amount of energy is radiated by antennas having the same current area. Therefore, it is convenient to chara-

cterize the radiation of an antenna by replacing it with some rectilinear equivalent antenna having the same current area. In the equivalent antenna, however, current is the same along its entire length and is equal to the current at a loop of a standing current wave in the real antenna (Fig. 6-9b). The length of such an equivalent antenna is called the *effective length* of the real antenna. This concept was introduced at the time when vertical antennas were mainly used. Therefore, this parameter is also referred to as *effective height*.

The more uniform the current distribution along the antenna wire, the closer is the effective height of the antenna to its physical height and the more effective the use of the radiating wire. The effective height of a half-wave dipole $h_{eff} = 0.64h_g$ where h_g is the dipole geometric length.

Introduction of this concept makes it possible to obtain the relationship between the intensity of the electric field produced by an antenna at some point of the space, the current, effective height and distance to the given point. This relationship is common for antennas of all types. The field intensity produced by an antenna is proportional to its current area, i.e. to the product of the loop current and the effective height. With the increasing distance from the antenna the field intensity decreases in inverse proportion to the distance, r , i.e.,

$$E = A \frac{I_{\max} A h_{eff}}{r} \quad (6-18)$$

where A is the proportionality factor depending upon the conditions of propagation of electromagnetic waves, the direction (with respect to the antenna) for which the field intensity is determined, and upon the selected system of units.

An electric field may be represented with the aid of vectors whose magnitude and direction display the magnitude and direction of the electric forces acting at different points

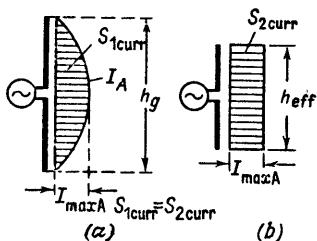


Fig. 6-9. To definition of the antenna effective height

(a) current distribution in actual antenna; (b) equivalent antenna with the same "current area" when current is distributed uniformly

in the space. An alternating current flowing through the wire of a rectilinear antenna produces in the surrounding space an electric field whose vectors are parallel to the antenna wire (Fig. 6-10). This means that an electric charge placed into this field moves parallel to the antenna wire. Such a field is generally known as the *field of polarized waves*.

Polarization is of great importance in radio engineering. For instance, if a horizontal receiving antenna is placed into

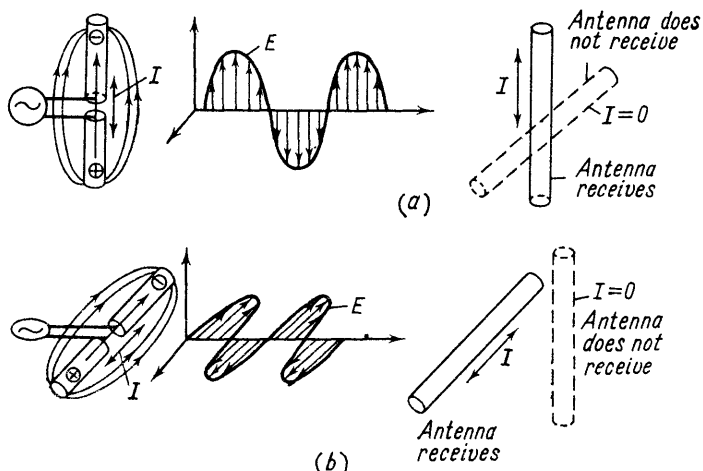


Fig. 6-10. Field polarization of linear radiator
(a) vertical polarization; (b) horizontal polarization

the field of vertically polarized waves, no current will be induced in it (Fig. 6-10a). The closer the direction of the receiving antenna axis to the direction of polarization, the greater the current induced in the antenna.

The intensity of the field radiated from a linear double-dipole is determined at a given point of the space by two factors: the polarization of radiation from the dipole and the summation of the waves radiated from different dipole sections. Let us consider these factors for the case of a half-wave dipole.

The electric field of the waves radiated in direction A (Fig. 6-11a) exhibits maximum intensity since its orientation coincides with the direction of the field produced by the so-

urce exciting the dipole. In addition, inphase currents flowing in both halves of the dipole produce inphase waves which reinforce one another. Therefore, radiation in direction A is a maximum and equals approximately the arithmetical sum of the waves radiated in this direction from separate dipole sections.

The electric-field intensity in the direction B (Fig. 6-11b) is much smaller since the electric field of wave E_{wave} is produced only by the dipole field component E_0 which is perpendicular to the propagation direction B . For the same reason, the dipole does not radiate along its axis since the field component perpendicular to the dipole axis is zero.

On the other hand, the waves radiated from dipole sections carrying inphase currents acquire an appreciable path difference Δ when propagating in direction B . Therefore, they will only partially reinforce one another depending upon the fact how much direction B differs from direction A .

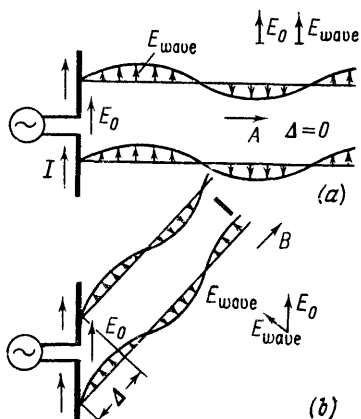


Fig. 6-11. Dipole radiation along different directions

To characterize the radiation of an antenna as a function of direction, *radiation patterns* are plotted. These represent in relative units the intensity of radiation from antenna along different directions lying in this or that plane. Planes in which radiation patterns are commonly given are the horizontal and vertical planes. Most frequently they are plotted in the polar coordinate system. In this case the distance from the centre to a given point on the curve drawn to a definite scale characterizes the antenna radiation intensity in this direction.

The radiation pattern of a half-wave dipole (Fig. 6-12) has its maximum at right angles to the dipole axis. Radiation intensity decreases with the decreasing angle and becomes zero along the dipole axis. Due to symmetry of the dipole,

radiation intensity is the same in all directions in the plane perpendicular to the dipole axis, and therefore the radiation pattern takes the form of a circle.

Directional radiation from antenna systems is usually desirable. For a rare exception, absolutely omnidirectional radiation is not only unnecessary but impracticable since

the directions in which any radiation is useless always exist and, vice versa, there are directions in which maximum-intensity radiation should be provided. It is common practice to design special antenna systems with highly pronounced directional properties. Generally, the radiation pattern is characterized by its width, by estimating it with the aid of an angle between the two directions in which the radiation intensity (power) is one-half its value in the main direction.

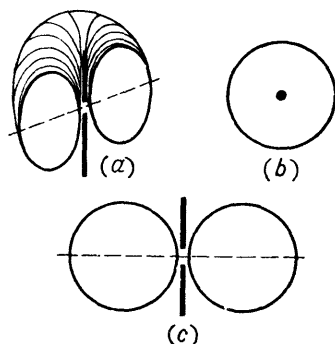


Fig. 6-12. Half-wave dipole radiation patterns

(a) space pattern; (b) radiation pattern in horizontal plane; (c) radiation pattern in vertical plane

Let us assume that a receiver receives signals from a transmitter supplied with an absolutely omnidirectional (isotropic) antenna which radiates energy uniformly in all directions. If we replace this antenna with a directional antenna oriented in such a way that its radiation maximum is directed towards the receiver, the volume of the received audio-signal will increase at the same radiation power. Consequently, such a replacement is equivalent to an increase in transmitter power. The number which indicates how many times transmitter power may be reduced due to the use of a directional antenna instead of an isotropic one, while maintaining the same strength of an antenna signal in the direction of its maximum radiation, is generally referred to as *directivity*. For instance, a half-wave dipole features a directivity value of 1.64.

Inasmuch as there are no absolutely isotropic antennas, a more realistic estimate of antenna directivity can be obtained by comparing the directivity of different antennas with

that of some frequently used antenna having a low directivity. A half-wave double dipole is generally used as the reference antenna. Therefore, in the further discussion the antenna directivity is defined with respect to the dipole radiation intensity in the direction of its maximum radiation. The product of this directivity value and the radiation efficiency of antenna is generally known as power *gain* (sometimes called gain) measured in decibels.

The foregoing refers primarily to transmitting antennas which radiate electromagnetic waves but the conclusions drawn also hold true for receiving antennas. For instance, if an antenna produces maximum radiation in some direction during transmission mode, then waves incoming from this direction to the same antenna being operated as a receiving one will induce maximum current in it. The waves incoming from those directions in which the antenna radiates no energy during transmission mode produce no effect on the antenna when it is used as a receiving one.

In other words, the directional properties of an antenna do not change, irrespective of whether it is used for transmission or reception. Therefore, when receiving signals from a given direction, it is always advantageous to use a directional receiving antenna which receives no interfering signals from radio stations located in other directions. It will produce a more powerful signal at the receiver input than an omnidirectional antenna. On this basis, we may say that the concepts of the directivity and power gain hold true when the antenna is used for reception. This property of *antenna reciprocity* makes it possible to confine our discussion to the characteristics of transmitting antennas without any loss in fuller understanding of antenna operation. There is, of course, a number of special receiving antennas. In majority of cases, however, these are used only in trunk networks and will not be discussed in this book.

6-3. EFFECT OF THE GROUND ON ANTENNA RADIATION. ASYMMETRICAL DIPOLES

So far we have discussed an antenna without any reference to the external factors which affect its radiating properties. The antenna operates, as a rule, near the ground influencing

both the radiation power of the antenna and its radiation pattern. In order to take into consideration the influence of the ground on antenna performance, let us first assume that the antenna is located above the surface of an ideal conductor. In this case, the electromagnetic waves radiated by the antenna arrive at some arbitrary point P in the space following

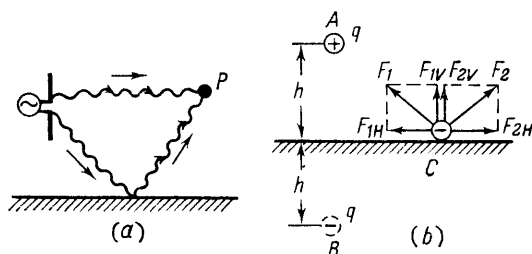


Fig. 6-13. Radiation from antenna placed above ideally conducting surface

(a) paths of direct and reflected waves; (b) electric charge and its image

two different paths: directly from the antenna and after being reflected from the conductor surface (Fig. 6-13a).

To find the resultant field at point P after summing the reflected and direct waves, it is convenient to use the image antenna concept which may be explained as follows. If the charge $+q$ (Fig. 6-13b) is placed at height h (at point A) above an ideal conducting surface, free charges on this surface are arranged under the action of electrostatic attraction so that the horizontal component of the electric field at the surface becomes zero. If it were nonzero, free charges might move along the surface under the action of this component. A stationary charge, however, cannot cause a continuous motion of charges in the conductor and, consequently, the electric field at the conductor surface must have a direction perpendicular to the surface.

A similar field is set up if the metal surface is removed and another charge equal in value and opposite in sign to the first one is placed at point B at a distance of $2h$ under the first charge. The second charge is called the "mirror image" of the first one or simply the "image charge". In this case, the electric field at arbitrary point C will be determined by

vectors F_1 and F_2 of the electric fields produced by charges $+q$ and $-q$. Inasmuch as charges are spaced equally from point C , vectors F_1 and F_2 are equal in magnitude. Horizontal components F_{1H} and F_{2H} of these vectors cancel each other, whereas vertical components F_{1V} and F_{2V} are summed. It is precisely this fact that makes it possible for the action of the conducting surface to be replaced by that of the image charge.

If the charge located above an ideally conducting surface starts to move vertically downwards, the image charge will also move to the surface, i.e. from the bottom upwards. If a

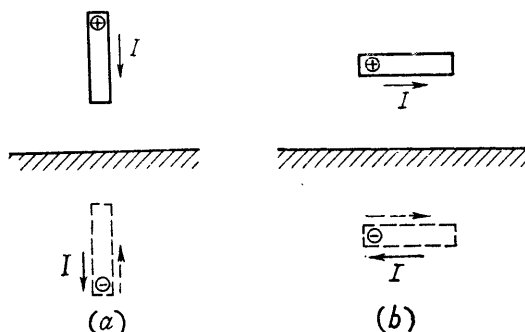


Fig. 6-14. Direction of currents in wires and in their images

wire is placed vertically above the surface, in which current at the given instant flows from the top downwards, then in the image wire charges of the opposite sign will move from the bottom upwards, i.e. the current in it will flow in the same direction as in the real wire (Fig. 6-14a). Therefore, if a vertical antenna is placed above the ideally conducting surface, its field can be determined as the resultant field produced by the real antenna and the image antenna in which the current of the same value and the same direction flows. Thus, the vertical antenna and its image form a system of two inphased dipoles.

If the charge placed above an ideally conducting ground moves horizontally, for instance, from left to right (Fig. 6-14b), the image charge having the opposite sign moves in the same direction. This means that if current in a horizontal

antenna flows in one direction, the current in the image antenna flows in the opposite direction. Thus, the horizontal antenna and its image form a system of two antiphase dipoles.

It is evident from the above that a perfectly conducting ground under an antenna changes basically its properties. If a vertical antenna is arranged so that its lower end touches the ground surface, then the combination of this antenna

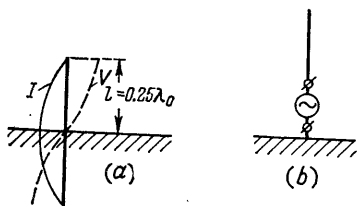


Fig. 6-15. Asymmetrical grounded dipole

(a) distribution of current and voltage in dipole and in its image; (b) way of connecting the driving generator

with the image antenna forms a double dipole (Fig. 6-15a) whose resonance wavelength is twice as long as the dipole length and, consequently, four times as great as the height of the antenna itself. Therefore, a grounded antenna is frequently referred to as a *quarter-wave dipole*.

Such a vertical aerial was suggested by A. S. Popov in 1895. Since then and up to the present day it is considered,

along with a double dipole, to be one of the main and most widely used antenna systems. The advantages of a grounded dipole are its simple construction and small length which are of special importance when operating at relatively long waves.

Inasmuch as the grounded dipole together with the image dipole form a half-wave double dipole, its radiation field above the ground surface is similar to that of a half-wave dipole (Fig. 6-16) whose lower-half radiation pattern is cut off by the ground surface.

Owing to the fact that the distribution of current and voltage along an asymmetrical grounded dipole is the same as for a double dipole, the nature of variations in its reactance and resistance is the same as that of the double dipole, and all the formulas derived for the double dipole hold true for the asymmetrical one. The only difference lies in the fact that the potential of the second terminal of the grounded generator is zero at any instant of time. Therefore, the potential difference between the terminal of the asymmetrical dipole and the ground is always twice as less as that between the

same terminal and the image dipole, i.e. than that of the double dipole. Consequently, the input resistance of the asymmetrical dipole proves to be twice as less than that of the double dipole, whereas the radiation resistance of a quarter-wave asymmetrical dipole is equal to one half of that of a

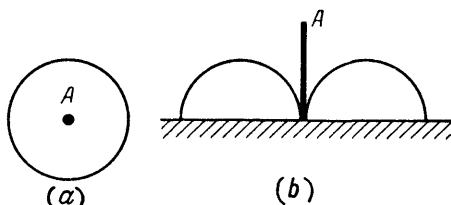


Fig. 6-16. Asymmetrical-dipole radiation patterns
(a) in horizontal plane; (b) in vertical plane

half-wave dipole, i.e. $R_{\Sigma}=36.6$ ohms. For the same reason the wave impedance of asymmetrical antennas may be considered to be equal to one half of that of symmetrical antennas, i.e. $Z_0 \approx 500$ ohms.

The foregoing holds true only when the ground presents an ideal conductor. If the ground exhibits poor conducting properties, the nature of current distribution inside the ground is changed, thus changing the radiation field of a dipole. Also, an increase in the resistance of the ground results in increasing losses in the entire radiating system (consisting of the dipole and the ground), decreasing current, decreasing antenna radiation power and efficiency, etc.

Ground resistance near the antenna base where all the currents induced by the antenna in the ground are concentrated is of particular importance. Ground metallization is generally used to improve the conductivity of this base area. For this purpose, metal sheets and wires are buried and the ground is impregnated with a variety of salts; sometimes antenna counterpoises are used (Fig. 6-17). Theoretical calculations show that the maximum losses occur in the zone having a radius of approximately 0.35λ . Therefore, ground metallization is accomplished in the radius of a half-wavelength.

Experience shows that there is no need to ground the antenna with a solid metal sheet. Sufficiently good results are generally obtained when a system of radially wires buried

at a depth of 20 to 50 cm is used. An increase in the total number of wires improves the quality of grounding but an increase in their number by more than 100-120 pieces has no appreciable effect. Grounding is improved when the radial wires are connected with jumpers.

Grounding is frequently accomplished with a system of wires elevated above the ground rather than buried. Such a system known as a *counterpoise* is designed to effectively screen an antenna wire from the ground, thus functioning as a perfectly conducting surface. Generally, the counterpoise gives worse results than antenna grounding but in

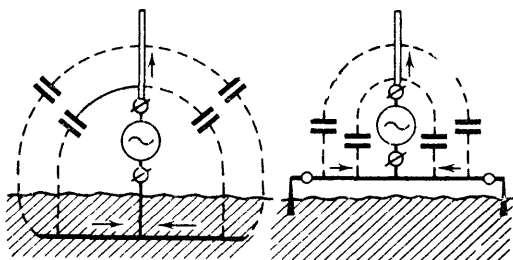


Fig. 6-17. Antenna grounding and counterpoise

some instances it proves to be technically suitable (for instance, in mobile radio stations, when installing a radio station on stony ground, etc.).

6-4. ANTENNA RESONANT FREQUENCIES. HARMONIC ANTENNAS

From resonant curves for a double dipole (Fig. 6-7), one may make sure that a single antenna possesses an infinite set of resonant frequencies. Resonance phenomena in an antenna occur each time when its length becomes equal to an integral multiple of half-wavelength as the wavelength is varied. The longest resonant wave is the one whose half length equals the antenna length. This wave is generally called the *dominant wave of the antenna*. The shorter waves for which the antenna proves to be tuned to resonance are known as *antenna higher harmonics*. To each such wave is assigned

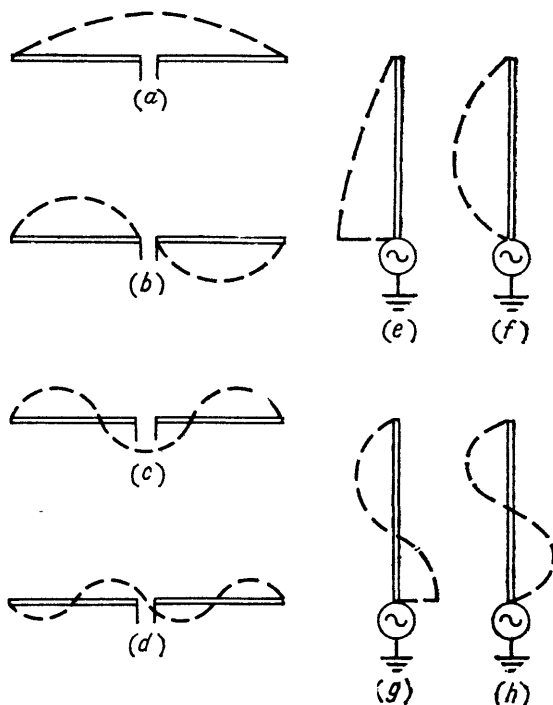


Fig. 6-18. Distribution of current in double (symmetrical) and asymmetrical dipoles at dominant wave and higher harmonics

a number corresponding to that of the half waves contained within the entire length of the antenna.

The validity of the above concepts is illustrated by the distribution of current standing waves in a double dipole whose length is equal to an integral number of a half-wavelength (Fig. 6-18).

Each half-wavelength section of the dipole wire can be considered as a portion insulated from others since the current at its ends is zero at any instant of time. As already stated, a half-wavelength wire represents a resonant system. Consequently, resonance takes place in the entire wire when its length becomes equal to an integral number of a half-wavelength. It is common practice to divide higher harmonics

into even and odd ones depending upon the fact whether an even or odd number of half waves is contained within the entire length of the wire.

Double dipoles are resonant systems which radiate particularly well when $l=n\lambda/2$, where l =the antenna length, λ =the wavelength, and n =any integral number. Asymmetrical antennas are also resonant systems which radiate parti-

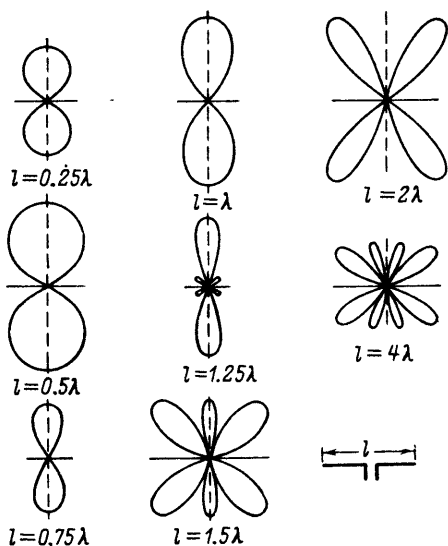


Fig. 6-19. Radiation patterns of double dipoles having different length

cularly well when $l=n\lambda/4$. The wave for which $l=\lambda/4$ is the antenna dominant wave.

When an antenna is operated at the dominant wave (Fig. 6-18a) or at all odd harmonics (Fig. 6-18c, e, g), its feed point is located at a current loop, and series resonance occurs in it. In these cases, the antenna input resistance is not high and is determined by the losses in the antenna circuit. At all even harmonics the antenna feed point is found to be located at a current node (Fig. 6-18b, d, f and h), and parallel resonance takes place in the antenna. In such cases, the antenna input resistance attains extremely high values.

When transferring from one odd harmonic to another odd harmonic or from one even harmonic to another even harmonic the input resistance of an antenna is comparatively little affected. On the contrary, the antenna input resistance changes significantly when transferring from an even harmonic to an odd one or vice versa. This fact makes it possible for antennas to be effectively operated at several harmonics (even or odd) without retuning. Antennas operating at higher harmonics are known as *harmonic antennas*.

In Fig. 6-19 the radiation patterns of double dipoles in the plane passing through their axis are shown for different values of the ratio l/λ ; the ground effects are ignored. To obtain radiation patterns in the vertical plane for asymmetrical grounded dipoles of twice less height, it is necessary to turn the patterns given through 90° and to cut off their lower halves along the dashed line.

From the above discussion it is clear that the input resistance and directional properties of a dipole depend upon the value of ratio l/λ or upon, as it often said, the "electrical length" of the dipole, i.e. its length expressed in fractions of the operating wavelength.

6-5. INPHASED AND ANTIPHASED ANTENNAS. REFLECTORS AND DIRECTORS

Directional radiation at short and ultrashort radio waves is frequently obtained with the aid of a system of dipoles arranged in a certain order relative to one another. It is common practice to use the dipoles so that currents in them are either in phase or antiphase. Therefore, such systems we shall call *inphased* or *antiphased*. The first antenna system of this type was developed in the USSR by M. A. Bonch-Bruyevich and V. V. Tatarinov.

Let us first consider the operating principles of antiphased antennas. A system consisting of two dipoles will radiate if the dipoles carrying antiphase currents (i.e. shifted by 180°) are located from one another at some distance D comparable with the operating wavelength (Fig. 6-20). Each such dipole radiates electromagnetic waves in all directions, which emanate from the dipole with phases determined by the phase of current in this dipole. The wave radiated by dipole 1 and

moving to the right will reach dipole 2 after a time interval $t=D/c$. If the distance D equals a half-wavelength, this wave will reach dipole 2 in a half-period after it has been radiated

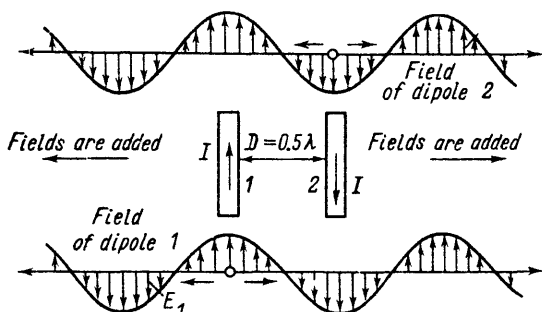


Fig. 6-20. Addition of the waves radiated by two antiphased dipoles

by dipole 1. During this half-period the current flowing through dipole 2, whose phase differs by 180° from that of the current in dipole 1 will change its phase by 180° and will radiate a wave with a phase equal to that of the wave coming from dipole 1. As a consequence, the two waves will add. It is quite evident that a similar picture will take place when waves move in the opposite direction, i.e. in the direction from dipole 2 to dipole 1.

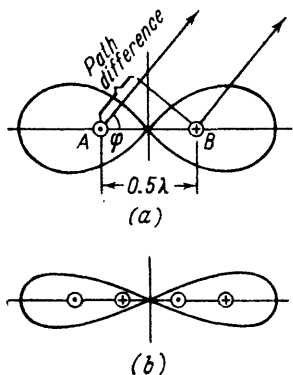


Fig. 6-21. Radiation patterns of antiphased dipoles in plane perpendicular to dipole axis (a) for two dipoles; (b) for four dipoles

In other directions (Fig. 6-21a) the path difference between both radiated waves is less since it is always less than the section AB . In these directions the waves are summed with different phases, and, therefore, the total field intensity is found to be less than in the direction of the system's axis (i.e. the line AB perpendicular to the dipole axes). In fact, when $\varphi > 45^\circ$ the two waves partially cancel each other. When $\varphi = 90^\circ$, i.e. in the

parallel to the dipole axes). In fact, when $\varphi > 45^\circ$ the two waves partially cancel each other. When $\varphi = 90^\circ$, i.e. in the

direction perpendicular to the system axis the path difference becomes zero and the waves cancel each other completely. Fig. 6-21a illustrates the radiation pattern of such an antiphased system in the plane perpendicular to the dipole axes. The radiation pattern in the plane passing through the dipole axes depends upon the fact whether each dipole is a double dipole or an asymmetrical radiator and upon its electrical length.

If another pair of such antiphased radiators is connected to the given system and arranged in the same plane so that the distance between the adjacent dipole wires is 0.5λ (Fig. 6-21b), then the intensity of radiation in the direction of the system's axis will increase. A very sharp radiation pattern whose maximum is directed along the antenna system axis can be obtained with a sufficient number of antiphased radiators. Such antennas fall into the category of *axial-mode radiators*.

It should be noted that the antiphased radiators must not necessarily be spaced at $\lambda/2$ intervals. Antiphased antennas with a lesser distance between the dipoles ($D < 0.5\lambda$) display a much higher directivity than the system discussed above.

This can be easily explained by considering the antenna radiation pattern when the distance between the dipoles $D = 0.5\lambda$ (Fig. 6-21a). The phase difference between the waves radiated by both dipoles in directions close to the axis of the system is small and the amplitude of the sum wave decreases comparatively little. On the contrary, the phase difference in directions forming a large angle with the axis of the system is large, and a small change in the direction causes an abrupt change of the sum-wave amplitude. Antennas with $D < 0.5\lambda$ are characterized by an appreciable phase difference even along the axis. Therefore, even a slight directional deviation results in an abrupt decrease of the sum-wave amplitude, and the radiation pattern proves to be much sharper than that of antiphased antennas having $D = 0.5\lambda$.

The lesser the distance between the dipoles in an antenna system, the higher is its directivity but the lesser is its radiation resistance. Therefore, it is common practice to take the distance D not less than 0.1λ .

Inphased systems, i.e. the systems consisting of the dipoles whose currents are in phase, have the radiation maximum in the direction perpendicular to the axis of the system since

the waves radiated by dipoles in this direction have no path difference and, consequently, are added up with the same phase (Fig. 6-22). In other directions, however, some path difference appears which increases with decreasing angle φ . The path difference reaches its maximum in the direction of the system's axis. In the case illustrated in Fig. 6-22a it equals a half-wavelength. Therefore, waves cancel each other in this direction. Thus, the inphased system provides maxi-

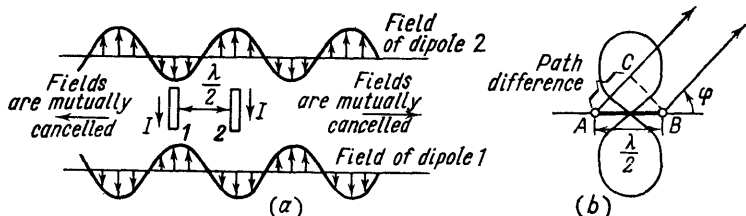


Fig. 6-22. Radiation of two inphased dipoles
(a) addition of the waves; (b) radiation pattern

imum radiation in the direction in which the antiphased system radiates no energy and does not radiate in the direction in which the antiphased system has maximum radiation.

Both inphased and antiphased antennas feature bi-directional radiation, i.e. they provide the identical maxima of radiation, in two opposite directions. Unidirectional radiation is obtained when currents in two dipoles spaced at a quarter-wave apart are shifted in phase relative to one another by a quarter-period. If current in dipole 1 (Fig. 6-23a) leads that in dipole 2 by 90° , then the wave propagating from dipole 1 towards dipole 2 reaches the latter after time $t = \lambda/4c = T/4$. By this moment of time the current in dipole 2 will have the same phase as the current in dipole 1 at the instant when the wave which now has reached dipole 2 was radiated. Therefore, the wave emanating from dipole 2 will be in phase with the wave which has come from dipole 1 and both waves will be summed up. As a consequence, the amplitude of the sum wave propagating in direction A will be twice as great as that of the waves radiated by each dipole.

The reverse picture takes place in the opposite direction B. A quarter-period passes until the wave radiated by dipole 2 the current in which lags by a quarter-period behind the

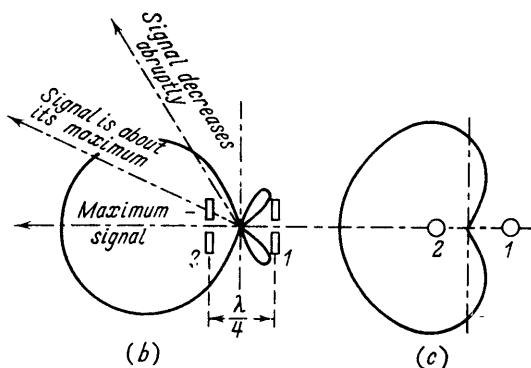
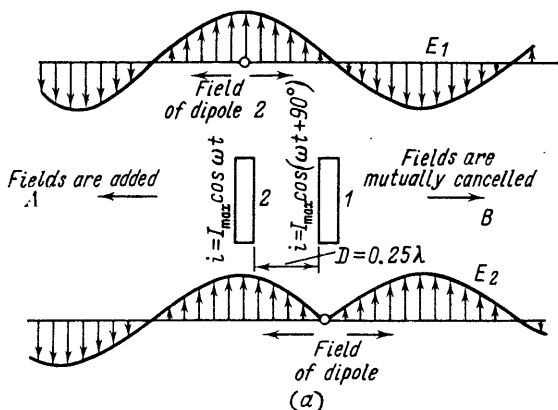


Fig. 6-23. Unidirectional antenna system consisting of two dipoles
 (a) addition of the dipole waves; (b) radiation pattern in dipole plane; (c) radiation pattern in plane perpendicular to dipoles

current in dipole 1 reaches the latter. By this time the current in dipole 1 will already lead by a half-period that in dipole 2 which gave rise to the wave approaching dipole 1. Thus, the wave indicated proves to be in antiphase with the wave radiated by dipole 1, and therefore, these waves cancel each other along the direction B. In any other direction the path difference between two waves has another value and the waves are either partially cancelled or partially summed up. As is apparent from the radiation pattern of such a

system (Fig. 6-23*b, c*), the dipole in which current leads in phase the current in another dipole acts in such a way as if it reflects the waves radiated by the latter sending them in the opposite direction. Owing to such an action, dipole 1 is called a *reflector* (or mirror) and dipole 2 is referred to as an antenna.

Unidirectional radiation can be also obtained at other distances between the dipoles but the phase difference between their currents should then have another value. If, for instance, the distance between the two dipoles is made less than a quarter-wavelength, the current through the reflector should lead that through the antenna by an angle exceeding 90° in order that the wave which has reached the reflector could meet the antiphased wave. In case of reduced distances between the antenna and the reflector some phase difference appears between the waves as they are summed up. Some phase difference is present even along the system's axis in the direction of maximum radiation. Therefore, broadside radiation from such antennas drops more abruptly than in the case of a quarter-wavelength distance between the dipoles, and their radiation pattern is more narrow.

The wave radiated from the antenna toward the reflector induces appreciable currents in it. If the induced current leads in phase the current through the antenna by 90° , the reflector performs its functions without requiring its own power source. The required phase shift can be always set up by appropriate reflector tuning which consists in selecting the length l of the reflector. The reflector may offer a resistance (when $l = \lambda/2$), a capacitive (when $l < \lambda/2$) or inductive (when $l > \lambda/2$) reactance to induced currents. As a result, the currents flowing

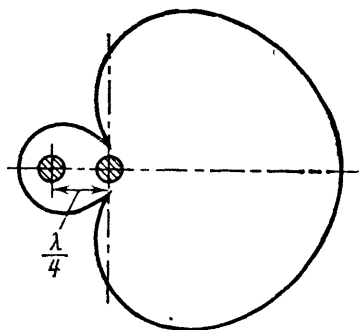


Fig. 6-24. Radiation pattern of antenna having passive reflector

through the reflector may be shifted in phase by one or another angle with respect to the wave which induces an electromotive force in it.

Radiation in the direction reverse to that of the maximum radiation cannot be compensated owing to the fact that the current induced in the reflector is always less than the current flowing through the antenna. Therefore, the radiation pattern of an antenna complete with such a reflector (Fig. 6-24) always proves to be somewhat worse than that of the antenna with a reflector which has its own power source.

The reflectors not requiring an independent power source and being excited by radiation from the antenna are called *passive reflectors* as distinct from those fed from a generator; the latter are referred to as *active reflectors*. The antennas with passive reflectors have a simpler feed system and are less complicated in tuning. They are, therefore, extensively applied.

In a system consisting of an antenna and an active reflector, the functions of the two dipoles may be interchanged by reversing, for instance, the direction of current in one of them. As this takes place, current in dipole 2 will lead that in dipole 1 and the antenna radiation pattern will change its orientation by 180° (Fig. 6-25). The same can be easily done in a system consisting of an antenna and a passive reflector.

The form and orientation of the radiation pattern of an antenna are determined by the distance between its dipoles as well as by a phase shift between the currents in the dipoles. A variety of radiation patterns may be obtained by selecting the appropriate distance and phase shift. Fig. 6-26 illustrates radiation patterns for an antenna with a passive dipole. Here each horizontal row shows the radiation patterns for a fixed distance between the antenna and the dipole as the tuning of the latter is varied. The tuning of the passive dipole is characterized by a phase angle θ_p between the current and voltage across it. The desired tuning is obtained by changing the length of the dipole or the length of a stub connected to it. By examining the vertical rows one may trace how the radiation pattern of the system varies at different distances D between the antenna and the passive dipole for a fixed tuning of the latter. It is easy to see that at a fixed distance D the

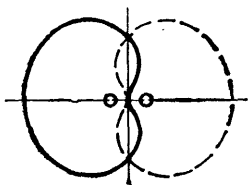


Fig. 6-25. Variation in radiation pattern of antenna with reflector when current phase in one of dipoles is changed by 180°

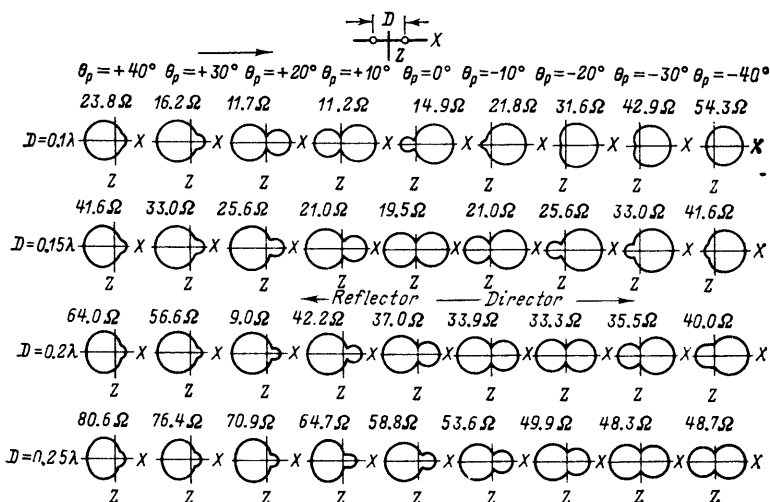


Fig. 6-26. Radiation patterns and radiation resistance of antenna with passive dipole for different phase angles and different distances between dipoles

passive dipole can exert no effect on the radiation pattern or may function as a reflector or, on the contrary, direct radiation towards itself depending on its tuning. In the latter case the passive dipole is referred to as a *director*.

From the radiation patterns shown in the vertical rows, one may make also sure that with a fixed phase angle θ_p the passive dipole can function as reflector or director depending upon the distance D between it and the antenna. Note that radiation in the direction reverse to that of the maximum radiation is compensated to a greater degree as distance D is decreased. This is explained by an increase of the current induced in the passive dipole as the latter is brought closer to the antenna.

Reflectors and directors, or in some cases both, are extensively used in antenna systems.

In Fig. 6-26 radiation resistance for the antenna system consisting of a half-wave active and passive dipole is also specified above each radiation pattern. As is apparent from the resistance values indicated, radiation resistance decreases with the decreasing distance between the dipoles.

6-6. GROUND EFFECT ON ANTENNA RADIATION PATTERNS

A horizontal wire arranged at a low height (as compared with a wavelength) above the ground radiates no energy as in this case it forms with the image wire a conventional two-wire transmission line. The distance between such an antenna and the image antenna increases with the increasing height and when it becomes comparable with the operating wavelength they form a radiating antiphased system. Naturally, the antenna radiation pattern varies depending upon the height above the ground. Fig. 6-27*a* illustrates the radiation patterns of a half-wave dipole located above the ground in the plane perpendicular to its axis. From these radiation patterns, one can see that a horizontal antenna located above a well conducting ground radiates no energy along the surface of the ground.

In the vertical plane passing through the dipole axis radiation intensity also varies with height. Radiation patterns in this plane can be determined with the aid of a general rule used to plot the patterns of complex systems. Most of such antenna systems include radiators which have identical radiation patterns and are oriented similarly in the space. This makes it possible to find the resultant radiation pattern of the system in two steps.

First the resultant radiation pattern of a complex system is calculated assuming that each its element radiates energy in all directions with some average intensity, then the irregularity of radiation of individual dipoles in different directions is taken into account by multiplying graphically the radiation pattern of the system of the omnidirectional radiators by the radiation pattern of the dipole (plotted in relative units). Fig. 6-27*b* illustrates the radiation patterns of a half-wave dipole located at height h , which were obtained by this means for the plane passing through its axis. For this purpose, the radiation patterns in Fig. 6-27*a* were multiplied graphically by the radiation pattern of a half-wave dipole plotted in the given plane in Fig. 6-12.

As can be seen from Fig. 6-27, the number of lobes of the radiation pattern as well as radiation directivity increase with the increasing height above the ground. In the direction

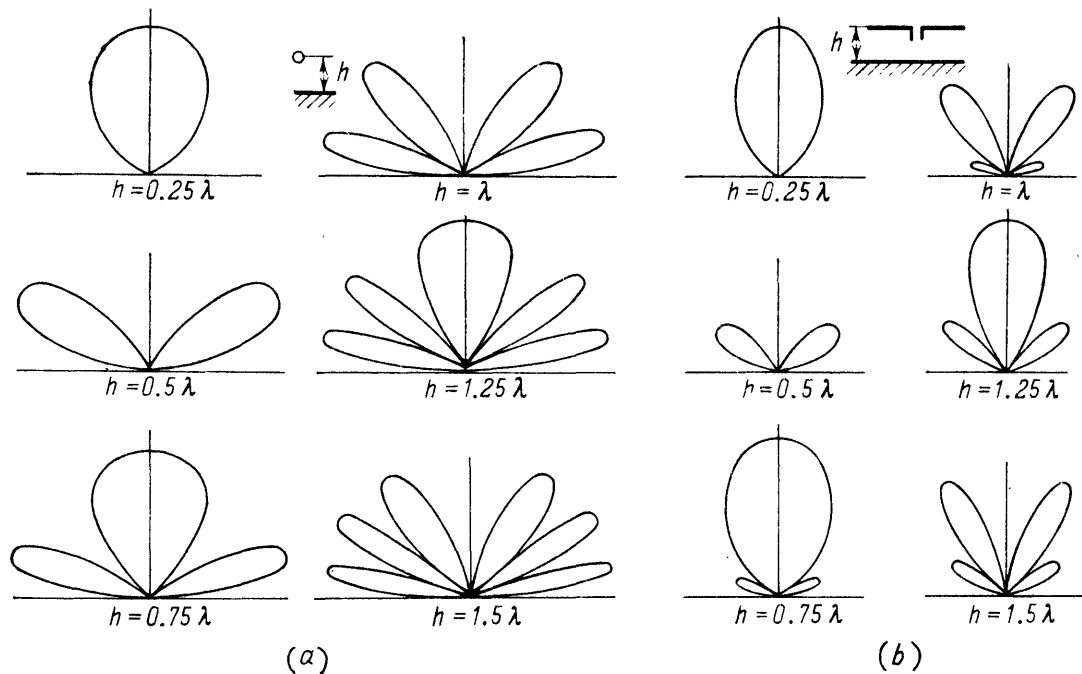


Fig. 6-27. Radiation patterns of half-wave dipole suspended at different heights above well conducting ground

of the maximum radiation the field of a horizontal antenna exceeds that of an omnidirectional radiator by 1.4-1.7 times as the height above the ground h varies from 0.25λ to 0.5λ , i.e. its directive gain is not great. It should be emphasized here that in the plane perpendicular to the dipole axis, in which the radiation pattern of the dipole has the form of a circle, radiation intensity increases at small angles to the horizon with the increasing height h . In this case, the greater h , the closer the lower lobe is to the horizon. As height h varies, radiation in the directions close to the vertical alternately appears and disappears.

As seen from the relationship between the radiation resistance of the horizontal half-wave dipole and its height above the ground (Fig. 6-28), radiation resistance increases rapidly with the increasing height, attaining at height $h=0.25\lambda$ the value of the dipole radiation resistance in free space. Radiation resistance is practically unaffected with a further increase in height. Consequently, approximate energy calculations for the dipoles suspended above the ground at a height exceeding 0.25λ may be accomplished without regard

to ground effects, especially as accurate evaluation of these effects is impaired by the difference in electric properties of soils. The radiation resistance of antennas having other dimensions varies in a similar way, hence in these calculations the plot shown in Fig. 6-28 may be used. Calculations may be carried out using the plot in Fig. 6-5 whenever radiation resistance increases or decreases in its absolute value.

The ground surface under the antenna does not display generally the properties of an ideal conductor and, therefore, alternating currents induced in the ground by the antenna electric field have an appreciable vertical component. Moreover, these currents produce an antenna radiation field. As a result, the field of an antenna suspended above the ground contains both horizontally and vertically polarized waves.

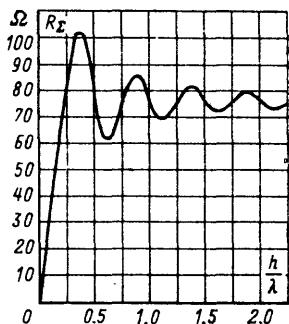


Fig. 6-28. Variation of radiation resistance of horizontal half-wave dipole as a function of its height above ground

Inasmuch as the ground presents a finite resistance to the currents induced in it, some energy radiated by the antenna is absorbed in the ground. The closer the antenna to the ground, the higher are the losses in the antenna and the lesser

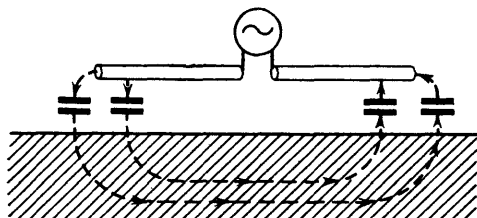


Fig. 6-29. Horizontal dipole located above poorly conducting ground

is its efficiency. As a result of absorption in the ground, the waves reflected by the ground are found to have lower intensity than in case of the reflection from an ideally conducting ground which absorbs no energy. Therefore, the resultant field over the ground surface differs from that over an ideal conductor. The major difference is observed when an antenna is

suspended at a low height. In this case a finite ground resistance may cause the radiation pattern of a half-wave dipole to turn through 90° . Thus, radiation in the horizontal plane proves to be directed along the dipole axis.

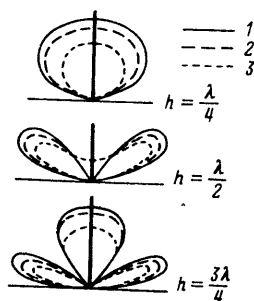


Fig. 6-30. Radiation patterns of horizontal dipoles located above surfaces having different conductivity; patterns are shown in plane perpendicular to dipole axis

This is explained by the fact that in case of a poor ground conductivity capacitive currents flowing through the dipole find their way into the ground (Fig. 6-29) and current paths form a loop whose radiation pattern in the horizontal plane takes the form of the figure eight extended along the dipole axis (see Section 6-8). In its turn, radiation produced by the current flowing through the horizontal dipole suspended at a low height is not intensive due to its image effect. This phenomenon manifests itself to a different extent at different antenna

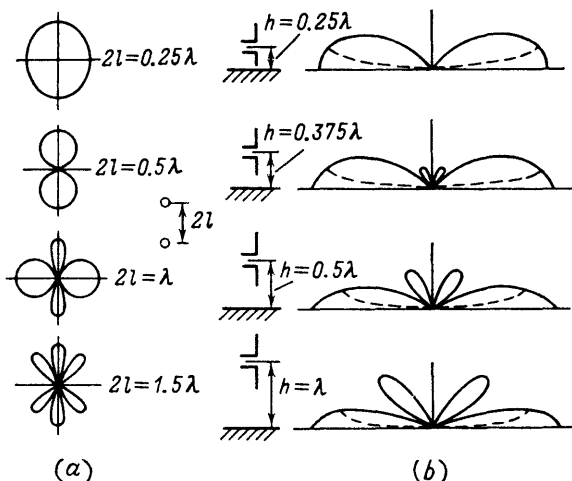


Fig. 6-31. Radiation patterns of vertical antennas located above ground (a) radiation patterns of two inphased omnidirectional radiators; (b) radiation patterns of half-wave double dipole (dash lines correspond to the case of poorly conducting ground)

heights above the ground and at different values of ground conductivity. In the general case radiation of the horizontal dipole in the direction of its axis is nonzero, which is an advantage when omnidirectional radiation is required.

Figure 6-30 illustrates experimental radiation patterns taken in the vertical plane for horizontal antennas operating at wavelength of 50 m and located above a well-conducting surface (1 — over sea water), semiconducting surface (2 — over wet ground) and above a poorly conducting surface (3 — over dry ground). These patterns indicate that the resultant field intensity is somewhat reduced with the deteriorating ground conductivity. These changes, however, are not very great. In practice, therefore, radiation patterns obtained for horizontal antennas over an ideally conducting ground are used.

A vertical antenna located above an ideally conducting ground forms with the image antenna an inphased system of dipoles arranged along a straight line. Such a system has a radiation maximum in the plane perpendicular to their axis,

i.e., in the horizontal plane and a radiation minimum in the vertical direction.

In Fig. 6-31a the radiation patterns of two inphased omnidirectional radiators are shown for different distances between them, and Fig. 6-31b presents the radiation patterns of a half-wave double dipole located at different heights above the ground. The number of radiation pattern lobes grows and radiation directivity along the directions close to the horizon increases with the height above the ground.

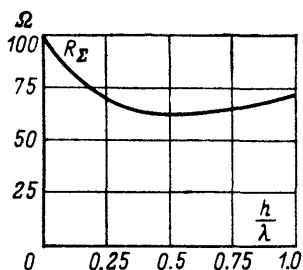


Fig. 6-32. Relationship between radiation resistance of vertical grounded half-wave dipole and its height above ground

At greater angles to the horizon new lobes appear; their intensity increases and they are lowered downwards with the height above the ground. Simultaneously, they become narrow and radiation directivity increases.

It is somewhat different with radiation from asymmetrical grounded antennas. Together with their images they form symmetrical antennas whose radiation patterns are determined completely by their length. Therefore, their radiation patterns may be

easily found using the appropriate radiation patterns of symmetrical (double dipole) antennas of different length given in Fig. 6-19.

Radiation resistance of vertical antennas is less than that of horizontal ones and depends generally upon the antenna height above the ground. At a minimum height when an antenna lower end is at the ground level, the radiation resistance of a half-wave dipole, as indicated above, is equal to 100 ohms. As the height of the half-wave dipole centre is raised above the ground, the radiation resistance quickly drops to that of the half-wave dipole in free space (73.1 ohms). Beginning from height $h=0.25\lambda$ (Fig. 6-32) all the energy calculations for vertical antennas may be performed without taking into account ground effects.

In the case of a poorly conducting ground the radiation pattern changes primarily at low angles to the horizon. Ground losses manifest themselves in the fact that the vertical antenna

ceases to radiate along the ground surface, and the lower lobe of its radiation pattern takes the form as shown in Fig. 6-31b by a dashed line. Simultaneously, a small additional lobe arises at a great angle to the horizon but its intensity is not great and usually it is disregarded. Naturally, the resultant radiation field intensity decreases due to ground losses as compared to an ideally conducting ground.

Similar changes occur in the radiation patterns of grounded dipoles (Fig. 6-33). In fact, these patterns indicate that a poor ground conductivity shows up primarily in the decreased field intensity and in the "repulsion" effect produced by the ground surface on the lower lobe of the radiation pattern.

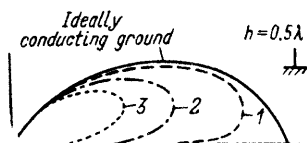


Fig. 6-33. Radiation pattern of vertical grounded dipole for different values of ground conductivity

6-7. COMPLEX DIPOLES

High radiation-resistance antennas feature high efficiency, low current and voltage at resonance and are simple in tuning and adjustment. Therefore, there is always a quest for the maximum increase of antenna radiation resistance by increasing the length of a radiating wire. For a linear wire, however, radiation resistance does not increase continuously with the length periodically decreasing by small values (see Fig. 6-5) due to radiation from the wire sections carrying antiphase currents.

An increase in radiation resistance through the use of the antenna longer wire can be achieved in two ways. First, the wire sections carrying opposite currents may be bent into nonradiating quarter-wave stubs as it is done in inphased antennas (see Fig. 6-1, *l*). In this case, however, the total wire length determining the resistance loss in the antenna system proves to be almost twice as great as the length of the radiating portion. Therefore, to obtain directional radiation, inphased antennas made of the wires arranged along one straight line are used. Second, an antenna wire may be bent at those points where the direction of current is changed. In this case, a spatial direction of currents in separate sections of such an

antenna system will not be reverse, and with a bending angle of 90° no compensation of radiation from these sections will occur. The system thus formed is referred to as a Chireix-Mesny antenna (Fig. 6-34).



Fig. 6-34. Chireix-Mesny antenna

If the antenna wire is bent through 180° , the spatial direction of currents in the antenna will be the same at all points. When the distance between the radiating sections is small compared to the wavelength, the waves produced by these sections reach any space point in the same phase and reinforce each other. Typical examples of the so-called *folded* or *bent dipoles* based on this principle are illustrated in Fig. 6-35. Any dipole of this type has a total length of wire equal to an integral number of the half-wavelength and, therefore, presents a resonance system whose input impedance is a pure resistance. Current supplied by a generator flows around the inphased radiating elements which are half-wave dipoles.

As distinct from inphased antennas consisting of dipoles arranged along a straight line, such systems do not contain nonradiating sections of wire and, therefore, resistance loss is reduced in them to a minimum. In return, the radiation pattern of folded dipoles does not differ practically from that of a conventional half-wave dipole.

Figure 6-35a illustrates a two-dipole folded antenna which is frequently referred to as a *folded-dipole antenna*. In this antenna the end A is connected with the end C so as not to disturb the system symmetry. Otherwise (Fig. 6-35b), different currents will flow through the feeder wires, which will cause intensive radiation in the surrounding space. It is necessary to say that the two dipoles are connected in series rather than in parallel as could seem at first sight. When the length of a folded antenna is equal to an odd number of the half-wavelength (Fig. 6-35c), the antenna system becomes symmetrical and there is no need to connect its ends. In Fig. 6-35d a four-element antenna is shown.

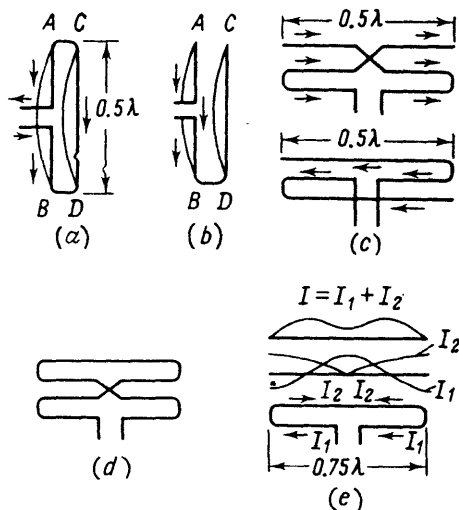


Fig. 6-35. Folded dipole antennas

(a) folded dipole; (b) folded dipole without short jumper; (c) three-element folded dipoles; (d) four-element folded dipole; (e) $3\lambda/4$ folded dipole

One can see from the above examples that identical inphase currents flow simultaneously through all wires which form a complex dipole. As these wires are located in the immediate vicinity from one another, such a system may be visualized as a single dipole with current $I_{\max A} = nI_{\max A1}$ flowing through it, where n = the number of elements in the antenna, and $I_{\max A1}$ = the current flowing through each antenna element.

Radiation power of a complex dipole can be determined from the following common formula:

$$P_{\Sigma} = \frac{1}{2} (nI_{\max A1})^2 R_{\Sigma1} = \frac{1}{2} I_{\max A1}^2 n^2 R_{\Sigma1}$$

where $R_{\Sigma1}$ is the radiation resistance of one element (of a half-wave dipole). Hence,

$$R_{\Sigma} = n^2 R_{\Sigma1} \quad (6-19)$$

i.e., the radiation resistance of a complex dipole is equal to that of one element multiplied by the square of the number of its elements.

The results of measuring the radiation resistance of dipoles agree well with the values obtained from formula (6-19). For instance, the measured radiation resistance R_{Σ} of a two-element folded dipole and a three-element dipole amounts to 320 and 640 ohms, respectively. Some discrepancy between the measurement results and values calculated by the formula (6-19) is explained by changes in current distribution in dipole wires due to their mutual influence. The radiation-resistance values indicated above are high enough for the two- and three-element dipoles to be matched directly with open (aerial) two-wire lines, this fact is very useful in applications.

Within any amateur's band complex dipoles perform sufficiently well without retuning the dipole and its feeder. In the case of a conventional dipole the feeder travelling-wave ratio may vary approximately two times within such a band. This property is also extremely valuable in wideband applications such as television, telecontrol and radiolocation where frequency-modulated signals are used.

A required radiation resistance may be obtained by bending the dipole wire not at the points where current changes its direction but at some other points along the wire. In this case, antiphase currents will flow through separate sections of the dipole and its radiation resistance will be somewhat reduced. For instance, a folded dipole consisting of two dipoles of 0.75λ long and having radiation resistance $R_{\Sigma} = 500$ ohms (Fig. 6-35e) is very suitable for use with a 500-ohm aerial line. The total wire length of such a dipole equals 1.5λ , therefore, it presents a resonance system. Current distribution in its two wires is different (see Fig. 6-35e). The radiation pattern of such a system differs little from that of a half-wave dipole. In some instances the radiation pattern changes simultaneously with the changing radiation resistance.

6-8. LOOP ANTENNAS

As distinct from the antennas discussed above, a loop antenna (Fig. 6-36a, b) is used at waves whose length is much greater than the natural wavelength of the antenna. A loop may be considered as a short-circuited line with its end sections spaced somewhat further apart. The loop natural wavelength is generally 4-6 times longer than the total length of

the wire constituting the antenna. The loop is tuned with the aid of capacitors or inductance coils connected at its input. Currents flowing through opposite sides of the loop antenna have opposite directions as its dimensions are usually much less than the operating wavelength.

The radiation pattern of a loop antenna in the plane perpendicular to the loop takes the form of a figure eight (Fig. 6-36c).

Antiphase waves radiated from opposite sides of the loop have no path difference in the direction perpendicular to the loop plane and, therefore, cancel each other. In other directions no complete cancellation occurs due to the presence of some path difference. Maximum path difference is observed in the direction of a straight line connecting the loop sides. In this direction radiation is a maximum.

Radiation resistance and effective height of such an antenna are small owing to the small length of loop sides (the distances between pairs of antiphased radiating elements are small compared to the wavelength). Although loop antennas are rarely used in transmitters, they are extensively employed as receiving antennas. In order to increase its effective height, the loop antenna may be made of several complete turns of wire. In this case, the shape of loop outlines may be rectangular or circular.

The effective height of a loop antenna is given by

$$h_e = 2\pi nS \frac{1}{\lambda}$$

where n is the number of wire turns which form the loop and S , the loop area.

The magnetic field and, consequently, the effective height of a loop may be increased if it is wound on a core fabricated from a material having high magnetic permeability and low

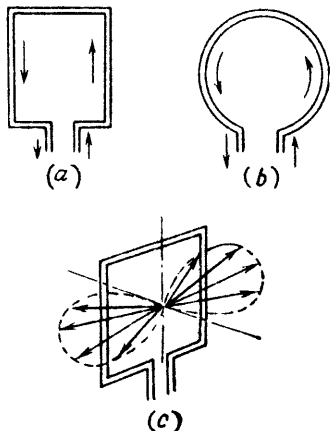


Fig. 6-36. Loop antennas
(a) square loop; (b) circular loop;
(c) radiation pattern

losses for high-frequency currents. Ferrite is most frequently used for this purpose. Antennas of this type have a wide application in receivers and are known as *magnetic* or *magnet-core antennas*.

6-9. LONG- AND MEDIUM-WAVE ANTENNAS

At long and medium waves the ground usually behaves as a good conductor. At the surface of such conductor the electric field may only be directed perpendicular to the surface. Therefore, both transmitting and receiving antennas designed for these waves must have a developed vertical portion. In

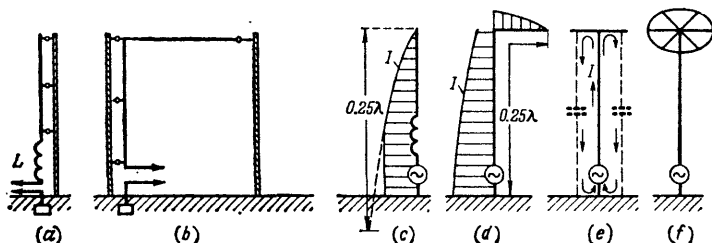


Fig. 6-37. Long- and medium-wave antennas

(a) grounded dipole with lengthening coil; (b) L-shaped antenna; (c) distribution of current in antenna with lengthening coil; (d) distribution of current in L-shaped antenna; (e) T-shaped antenna; (f) antenna with spiderweb hat

order that an antenna be at resonance and have a sufficiently high radiation resistance as well as high efficiency, its size should be close to at least 0.25λ . Thus, at long waves, antennas with a height equal to several hundreds of metres are required. In practice, however, antennas (masts) with a height not exceeding 200-300 m can be constructed. Therefore, at waves longer than 1000 m one is generally forced to use antennas whose length is less than the resonant one. As a result, antenna input impedance has, in this case, a reactive component of a capacitive type which can be compensated by an inductance coil connected in series with the antenna (Fig. 6-37a). Such coils are known as *lengthening coils*. As can be seen from the plot shown in Fig. 6-5, the radiation resistance of antennas having a small electric length is extremely low. On the other hand, the resistance of lengthening coils

is rather high. Therefore, the resistance loss in the antenna circuit becomes greater than or equal to the radiation resistance, and the antenna efficiency, according to formula (6-7), is found to be extremely low.

When a medium-wave antenna operates in a wide range of frequencies, the frequency of the signal supplied to it may prove to be less than the resonant frequency. In this case, the reactive component of its input impedance is inductive and a capacitor known as a *shortening capacitor* is used to tune the antenna. In the general case, the tuning circuit of a wideband antenna should contain both a capacitor and inductance coil. From the viewpoint of obtaining antenna maximum efficiency, it is advantageous, of course, to use a shortening capacitor but this is not always possible because of the difficulties associated with the construction of a long antenna.

Application of tuning elements does not change the antenna radiation resistance which is determined only by the antenna electrical length and, therefore, short antennas always have a low radiation resistance. Thus, to obtain high radiation power from short antennas, they must be excited with high currents. A low radiation resistance also makes the antenna resonance curve very sharp. This means that the antenna is very critical in tuning. In addition, for an antenna having a low radiation resistance its lower end must be thoroughly grounded, otherwise, the antenna efficiency will be drastically reduced.

To increase efficiency, instead of employing an inductance coil, the antenna length is frequently increased up to the resonant one. In this case the antenna wire is bent at a right angle at the height of the supporting mast so that the remaining section of the wire forms a horizontal portion. Such an L-shaped antenna radiates better than a straight antenna fitted with a lengthening coil, but it requires a second mast to be installed (Fig. 6-37b). If the L-shaped antenna is suspended at a low height, its horizontal portion hardly radiates since it forms with its image a two-wire line, but current distribution in the radiating vertical portion is improved significantly. A part of a current standing wave near the current loop is now contained within the antenna vertical portion, with the loop being located closer to the antenna upper end

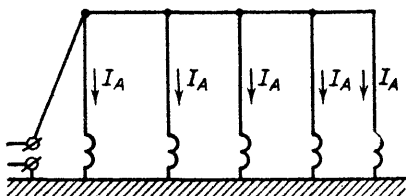


Fig. 6-38. Long-wave antenna consisting of several closely spaced radiators

which is now under the most favourable conditions for radiation.

The current at the antenna upper end may also be increased if the antenna horizontal portion is made in the form of two

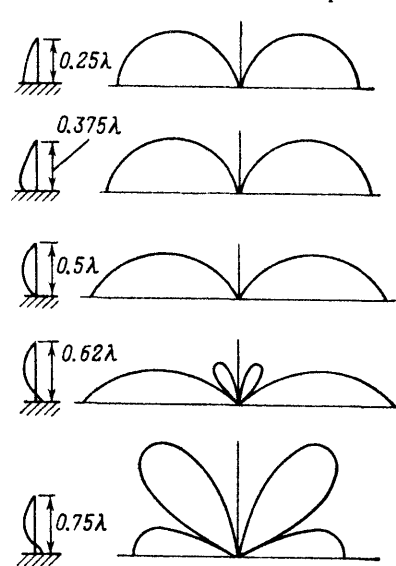


Fig. 6-39. Vertical radiation patterns of grounded dipoles having different length

horizontal arms (the T-shaped antenna shown in Fig. 6-37e) or in the form of several arms as in Fig. 6-37f. In all cases horizontal elements and the ground form some capacitance. Owing to this, the current is not already zero at the end of the antenna vertical portion, and the current distribution along this portion becomes more uniform. As a result, current area and, consequently, the effective height of the antenna are increased.

To increase radiation resistance, broadcast-band antennas are frequently made in the form of several vertical dipoles fed from one generator. Dipoles are spaced at the distances which

are much less than the wavelength and are connected together with upper horizontal portions (Fig. 6-38). Such a system may be considered as a combination of T- and L-shaped antennas.

From the comparison of the radiation patterns of grounded dipoles having different length (Fig. 6-39), it is obvious that the dipoles with electrical length l/λ ranging from 0.5 to 0.62 display the maximum radiation at small angles to the horizon (it is very important for long-range communication). Research has shown that the most advantageous antenna length $= 0.528\lambda$. Antennas having such a length are quite feasible in the medium wave band.

6-10. SHORT-WAVE ANTENNAS

The shorter the wave, the greater is the variety of antenna types used. At short waves ground conductivity deteriorates, thus increasing the losses in an antenna grounding. Therefore, the use of grounded dipoles is generally avoided in this wave band. Grounded dipoles yield good results only in the vicinity of large water surfaces or when the radio stations are located on wet soils. The best results are obtained with half-wave dipoles. In this case a current loop is at the grounding point, and current losses will be low even when the contact resistance is high. If it is desirable to reduce the mast height, the dipole may be bent at a right angle to form an L-type antenna or to use an inclined one shown in Fig. 6-40.

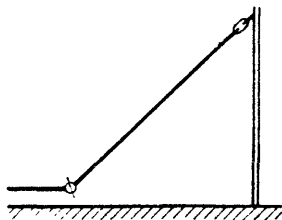


Fig. 6-40. Inclined wire

A horizontal half-wave dipole is the most widely used type of short-wave antennas. To obtain the maximum range of communication, it is desirable to have the antenna intensively radiating at low angles ($10-20^\circ$) to the horizon. From this point of view horizontal dipoles provide good radiation patterns when being suspended at heights varying from 0.5λ and more (see Fig. 6-27a). It is usually very difficult, however, to increase the suspension height greater than $h=\lambda$.

In the horizontal plane a short-wave antenna has an irregular radiation pattern; in a plane passing at an angle φ to the horizon, the magnitude of this irregularity changes depending upon the antenna height above the ground and upon the angle φ (Fig. 6-41). This should be taken into account during the

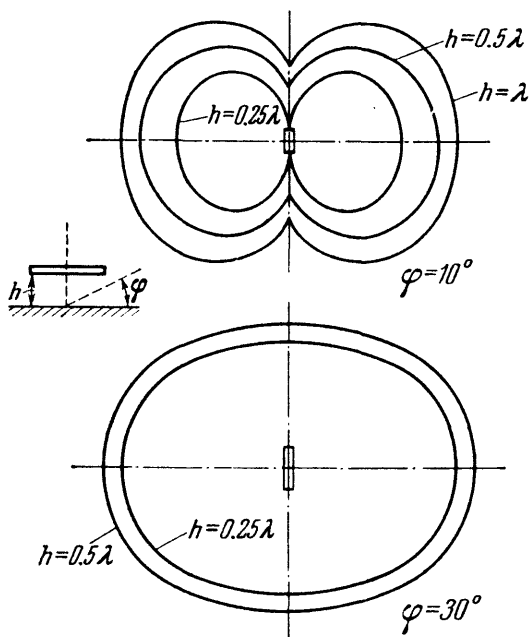


Fig. 6-41. Horizontal radiation patterns of horizontal half-wave dipole for different angles to horizon and different heights above ground

installation of the antenna by orienting the radiation maxima in the direction of the most important communication channels. The simplest method of feeding the half-wave dipole, which gives good results in the desired wave band, is to feed the dipole at its centre through a two-wire aerial feeder.

As the dipole input impedance is close to 73 ohms and the feeder offers a wave impedance of some hundreds of ohms, the dipole and feeder will be mismatched and standing waves will arise in the latter if no special precautions are taken. Among disadvantages of this feeding method are high voltage at some sections of the feeder and its low efficiency due to the presence of standing waves as well as significant variations of feeder input impedance along its length. These effects are tolerable only when the feeder length does not exceed 3λ - 4λ . In some instances, however, even shorter feeders must be matched.

Sometimes, design features require that the feeder be connected to one of the ends of the half-wave dipole. As a current loop is located at the end of the dipole, the free end of the dipole second wire is attached to an insulator (Fig. 6-42) to retain feeder symmetry, i.e., to obtain a current loop at this end.

If the length of the feeder exceeds 3λ - 4λ and its wave impedance differs from that of the antenna, losses in the feeder become too high and care should be taken of matching the feeder to the antenna. For this purpose matching transformers are generally used. Fig. 6-43a shows how a high-resistance aerial feeder can be matched to a low input-impedance dipole with the aid of a transformer made in the form of a quarter-wave open-circuited line. The match is achieved by changing the position along the transformer line where the feeder is connected to it. The closer the connection points to the open-circuited

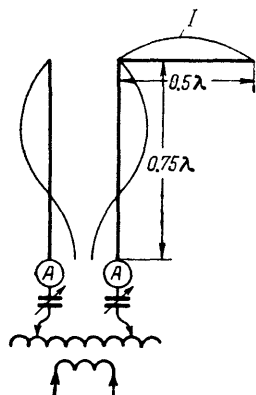


Fig. 6-42. Connection of feeder to half-wave horizontal dipole

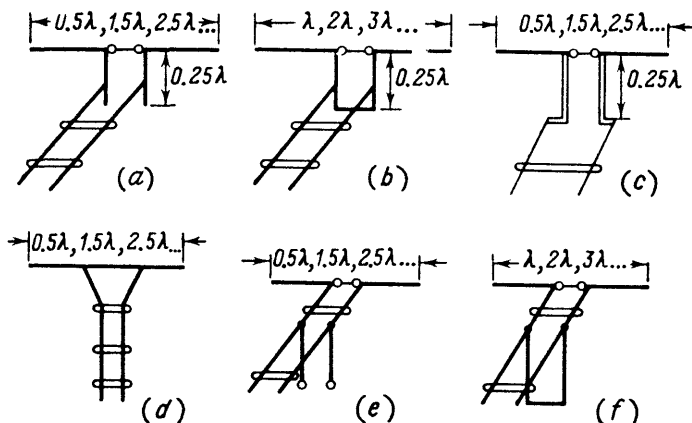


Fig. 6-43. Matching the horizontal dipoles to two-wire feeders

end of the transformer line, the higher the input impedance of an antenna-transformer system.

Figure 6-43b illustrates the method of matching high input impedance dipoles to an aerial line. In this case, a section of a short-circuited line is used as a matching transformer.

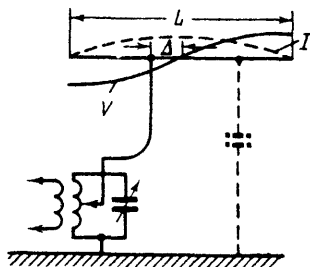


Fig. 6-44. Matching the horizontal dipole to single wire feeder

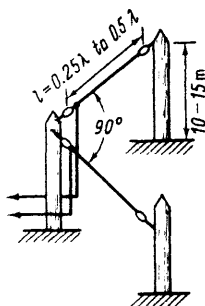


Fig. 6-45. Pistolkor's V-antenna

In the system shown in Fig. 6-43c the dipole input impedance is matched by selecting the required wave impedance of an intermediate quarter-wave line. A matching technique using a reactive stub made in the form of a section of an open- and short-circuited line is illustrated in Figs. 6-43e, f. Here, the match is obtained by selecting the stub length and the position where it is connected to the feeder.

Another technique of matching half-wave dipoles with the aid of a section of a divergent two-wire line (Fig. 6-43d) is very widely applied in antenna systems. On the one hand, a match is obtained here by changing the position where the line is connected to the dipole and, on the other hand, by changing the wave impedance of the divergent section. The closer the connection points to the dipole ends, the greater the input impedance of the dipole. The same principle is used in matching dipoles to a single-wire feeder (Fig. 6-44) by selecting the feeder connection point with respect to the dipole centre; the selected position is characterized by displacement Δ . In this case, radiation from the single-wire feeder is low due to the fact that under matched conditions almost all energy is transmitted along the line (feeder) and absorbed by the load.

Pistolkors' V-antenna (Fig. 6-45) has found extensive application in short-wave radio stations. It presents a double dipole whose halves are spaced through 90° (sometimes 120°) rather than through 180° . The radiation resistance of such an antenna is somewhat reduced, but its radiation pattern

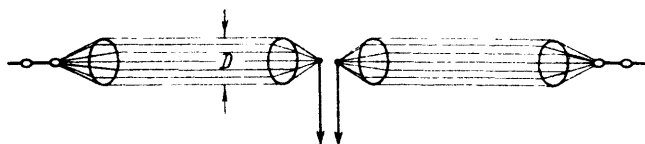


Fig. 6-46. Nadenenko dipole

in the horizontal plane is more uniform than when the dipole wires are arranged along one straight line. This antenna is practically omnidirectional.

In order that an antenna perform well in a wide band of frequencies, its radiation pattern and input impedance must remain relatively constant as the wavelength is changed. This property is inherent in the low impedance antennas consisting of large-diameter dipoles. The widely used Nadenenko dipole, also called a cage antenna (Fig. 6-46), is based on this principle. With radiator diameter $D=0.03\lambda$, the dipole wave impedance is close to 300 ohms and the antenna can be effectively matched to a 200-ohm feeder (usually a four-wire line) in a wide band of frequencies. In this case, there is no need to fabricate the antenna from a solid conductor (experience shows that six or eight wires arranged so as to form a cylinder can successfully replace a solid conductor of the same form).

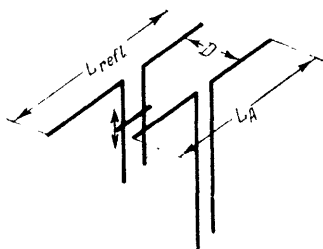


Fig. 6-47. Horizontal half-wave dipole with passive reflector

Owing to relatively small length of dipoles, good possibilities exist for designing short-wave directional antennas. The antenna fitted with a reflector is the simplest one among these, whereas antennas complete with a reflector and director are extremely bulky and therefore are rarely used. For

the standard distance between the antenna and the passive reflector, the sizes of such a two-dipole system shown in Fig. 6-47 are calculated by the following formulas:

$$L_A = \frac{143.4}{f}; L_{refl} = \frac{146.6}{f}; D = \frac{75}{f} \quad (6-20)$$

Here, L_A , L_{refl} and D are given in metres, and f in megacycles. For accurate tuning of the passive dipole, reactive stubs suggested by V.V. Tatarinov are frequently used.

At short waves the quarter-wave distance between the dipoles turns to be rather considerable, therefore, reduced distances between the dipoles are frequently used. It is not possible to specify accurately the most optimum dimensions of such a system because of the interaction between the dipoles. These dimensions also depend upon the design features of the system such as the antenna height above the ground, ground properties, wire diameter, etc. It is common

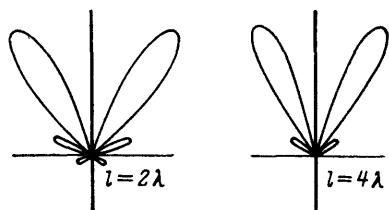


Fig. 6-48. Radiation patterns of travelling-wave wire

practice to determine these dimensions experimentally when adjusting the antenna. With reduced distances between the dipoles three-dipole antennas are feasible. In these antennas, the distance between the active dipole and the reflector is usually equal to 0.15λ , and that between the active dipole and the director, 0.1λ . The directivity of such a system, when tuned correctly, is increased by 10 dB and its radiation in the opposite direction is reduced by 30 dB.

In the short-wave range travelling-wave antennas are extensively used. To produce a travelling wave in antenna wires a pure resistance equal to the antenna wave impedance is connected at their ends. The radiation pattern of a travelling-wave wire has the form of a narrow lobe (or more accurately a body obtained by rotating this lobe about the wire as about an axis) oriented in the direction of wave motion (Fig. 6-48). Radiation patterns become more sharp and the angle of lobe inclination increases with the increasing electrical length of the wires.

Antenna directivity may be increased by using two (V-antennas) or four (rhombic antennas) wires arranged at some angle to one another so that the direction of lobes in their radiation patterns would coincide (Fig. 6-49). In rhombic antennas a lossy line (for instance, made of steel wires) is generally used as the load resistor. The side length of travelling-wave antennas is usually selected between 2λ and 4λ . The mast height of a V-antenna and the angle between the sides

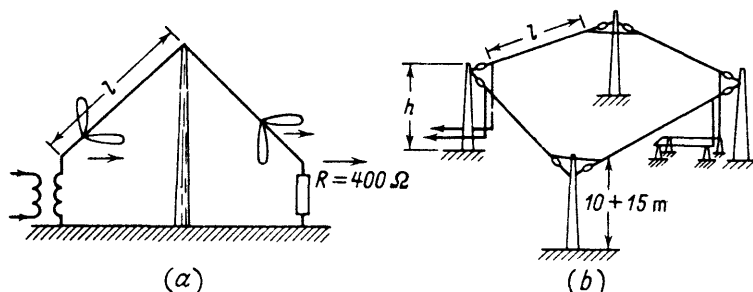


Fig. 6-49. Travelling-wave antennas

(a) V-antenna; (b) rhombic antenna

of a rhombic antenna must be selected so that the wire main lobes have the desired direction. The major advantage of such antennas is their very wide band of operating frequencies, however, considerable dimensions and high power dissipated in the load resistor are serious disadvantages.

Inphased and antiphased antennas are also applied in the short-wave range. These antennas must perform two main functions: to produce radiation at low angles to the horizon and concentrate the radiated energy in the desired direction. To meet the first requirement in inphased antennas two, four and more dipoles are stacked at $\lambda/2$ intervals one above the other (Fig. 6-50). In this case, dipole radiation is summed up in the horizontal direction and cancelled in the vertical direction; this is explained by the $\lambda/2$ path difference. To obtain the same effect in antiphased antennas, dipoles must be located at one level at $\lambda/2$ intervals from one another (Fig. 6-51). Again, owing to the $\lambda/2$ path difference, the waves radiated by antiphased dipoles will reinforce or cancel each other in the horizontal and vertical directions, respectively.

To satisfy the second requirement, antenna arrays consisting of horizontal rows of dipoles fed in phase are used. The greater the number of dipoles in each row, the higher the antenna directivity in the horizontal plane. To provide inphase currents in the dipoles, they are connected to one feeder at

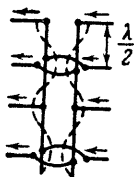


Fig. 6-50. Four-dipole inphased antenna

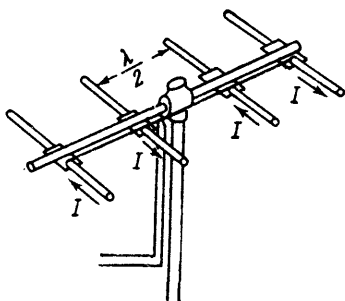


Fig. 6-51. Directional antiphased antenna

$\lambda/2$ intervals, with alternate vertical rows of dipoles being connected to the feeder by cross jumpers.

There is another widely used method of feeding dipoles in phase when they are connected in series through nonradiating phasing stubs (Fig. 6-52). To meet the condition of in-

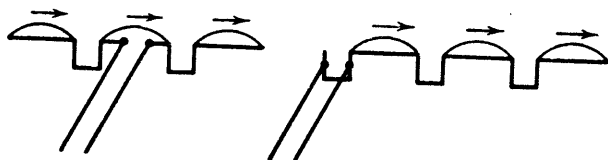


Fig. 6-52. Feeding the inphased dipoles through nonradiating phasing stubs

phase radiation, the total wire length of a dipole and the adjoining stub must be equal to the wavelength while the length of the dipoles themselves in this case is of no importance. Inphased antennas offer a high input impedance approximately equal to the sum of radiation resistances of all the dipoles and, therefore, to feed them high-impedance aerial feeders are most suitable.

To improve directivity and reduce antenna sizes, the distance between the dipoles in antiphased antenna is generally reduced down to 0.15λ - 0.1λ . This also simplifies the antenna feeding system since instead of phasing stubs, short cross jumpers connecting the dipoles to the feeder can be used (Fig. 6-53). Interaction between closely spaced dipoles may

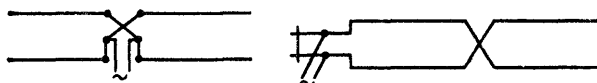


Fig. 6-53. Antiphased antennas with reduced distances between radiators

significantly affect their natural wavelength; it is, therefore, found experimentally during antenna adjustment. When dipoles are closely spaced, their radiation resistance drops drastically, thus, to feed them through a high-impedance feeder, a matching transformer must be used.

6-11. ULTRASHORT-WAVE ANTENNAS

In the ultrashort-wave range highly directional antennas (with high directivity at least in one plane) are primarily used. For the small wavelengths these antennas are so compact that no problem arises to make them rotatable. Thanks to this possibility, radio signals can be transmitted (or received) in any desired direction, thus, obtaining considerable economy in power and reducing radio station interference. The multidipole inphased and antiphased systems described above are most frequently used in the meter wave range.

At decimeter and centimeter wavelengths *reflectors* of different types, particularly *parabolic mirrors* made usually of duralumin sheets, are widely used to provide high antenna directivity. The principle of their operation is based on the known reflecting property of parabolic mirrors, which consists in the fact that the incident rays parallel to the mirror's axis converge (after being reflected) at one point located in front of the mirror and called a *reflector focus*. On the contrary, if a radiator is placed at the mirror focus F , the incident rays will be formed by the reflector into a parallel beam (Fig. 6-54). This occurs, however, only for an ideal point radiator. Real radiators present generally half-wave dipoles

being fed through a coaxial cable and a balancing quarter-wave transformer which isolates one dipole half from the cable braid (Fig. 6-55a). In order to make the radiation pattern as narrow as possible and to eliminate its side lobes, the size of the reflector (its diameter D) must be many times as great as that of the radiator, l , i.e., $D \gg l = \lambda/2$.

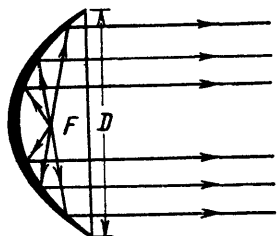


Fig. 6-54. Parabolic reflector focusing the radiation from point radiator into parallel beam

To obtain a narrow radiation pattern, the parabolic mirror must be machined with high precision.

Directional antennas are characterized by a beamwidth θ which can be determined from the formula

$$\theta = 60 \frac{\lambda}{D} \quad (6-21)$$

This formula indicates that with a mirror diameter being great enough an extremely narrow beamwidth can be obtained. For instance, with $D = 20\lambda$, the beamwidth $\theta = 3^\circ$, which is quite feasible in centimeter and even decimeter wave ranges.

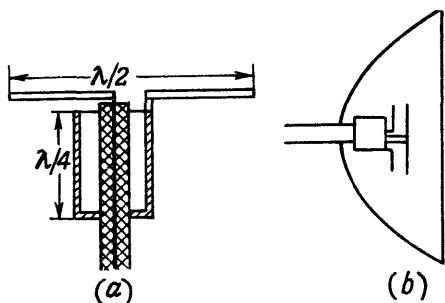


Fig. 6-55. Parabolic-antenna dipole (radiator)

(a) connection of dipole to coaxial cable; (b) dipole together with counter-reflector

Irregularities in the radiation pattern also occur due to the fact that not all radiation from the dipole comes to the reflector, since nearly half of its energy is radiated in the opposite directions from such a low directional dipole. This disadvantage is frequently eliminated with the aid of a *counter-reflector* placed ahead of the dipole (Fig. 6-55b).

In the case of metal dipoles which we will refer to as electric radiators, radiation is caused by the motion of electric charges and their electric field which is directed along the dipole and changes as a sine wave. If we build a device having the form of the same dipole, whose volume is partially "filled" with an alternating magnetic field changing by the same sine law, then the resultant "magnetic dipole" will radiate energy similarly to the electric one except for the fact that the electric field will be replaced with the magnetic one and vice versa.

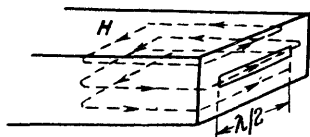


Fig. 6-56. Half-wave slot antenna (magnetic dipole)

A half-wavelength slot cut in the end face of a rectangular waveguide parallel to its wide wall can serve as such a device (Fig. 6-56). In this device the magnetic lines of force are directed along the slot in the same way as the electric lines of force are directed along the metal dipole. Therefore, such a radiator may be called a magnetic dipole. Its resonant and radiating properties will be fully identical to those of the metal dipole of the same shape when replacing the magnetic field with an electric one and vice versa. For instance, to tune the magnetic dipole to resonance it should be shortened somewhat, just like the metal dipole. To obtain the maximum bandwidth, the slot should be as wide as possible (with its length being further shortened). The maximum bandwidth is provided when the slot is cut along the entire waveguide width. In this case, to tune the dipole, the slot must be shortened appreciably, i.e., the waveguide end face must be partially closed (Fig. 6-57). The use of slots cut into metal surfaces as radiators was first suggested by M.S. Neumann who called them diffraction radiators. At present, radiators of this type are widely applied in the centimeter wave range.

From the above it follows that the open-ended waveguide should possess low directivity as it acts like a linear dipole. Its directivity, however, can be increased with the aid of a horn fitted to the open end of the waveguide (Fig. 6-58a). The wider the horn aperture and the greater the horn length, the smoother the transition from the waveguide to open space and the narrower its beamwidth. The horn aperture presents a system of inphased magnetic dipoles arranged in one line

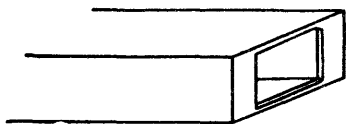


Fig. 6-57. Waveguide with end face partially closed

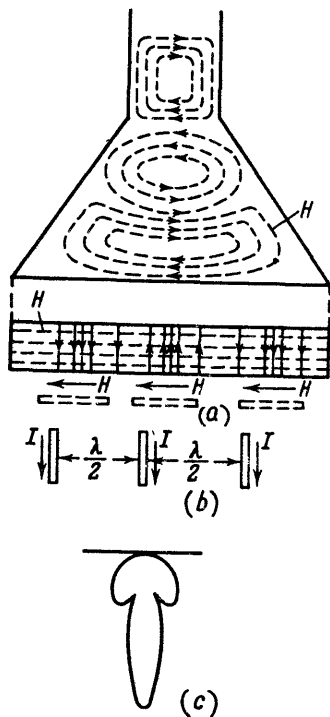


Fig. 6-58. Horn antenna

(a) construction; (b) horn aperture as a system of inphased magnetic dipoles; (c) radiation pattern

(Fig. 6-58b), which is equivalent in its radiating properties to the same number of inphased electric dipoles located at right angles to the magnetic ones. Therefore, the horn has a sharp radiation maximum (main lobe) in the direction of its axis and a few side lobes (Fig. 6-58c).

As already said, to obtain high directivity, the length of the horn should be much greater than that of the waves. For instance, to obtain a beamwidth of 50° , the horn length should be 8λ - 10λ . To narrow the beamwidth two times, the horn length must be increased 5 times. This, of course, can be accomplished only at centimetre and millimetre waves. The greater the horn length, the smaller may be its flare angle in order to obtain the prescribed aperture and, consequently, the directive gain.

The horn illustrated in Fig. 6-58a is referred to as a sectoral one since it is flared in one direction only. Horns of this type provide directional radiation in one plane only. To narrow the beamwidth in the plane normal to the horn large dimension, the horn should also be flared in this plane. The

resultant pyramidal horn (Fig. 6-59) essentially presents a stack array of inphased magnetic dipoles having a beam narrowed in two directions. When round waveguides are used, the horn takes the form of a truncated cone (Fig. 6-60).

In some applications it is desirable to provide high directivity in the vertical plane with the radiation maximum being directed horizontally (which is particularly important

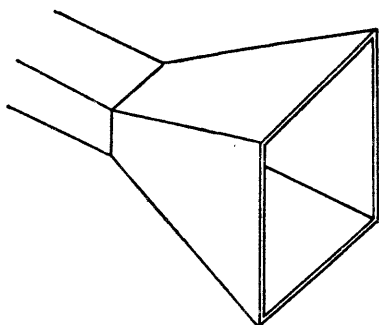


Fig. 6-59. Pyramidal horn

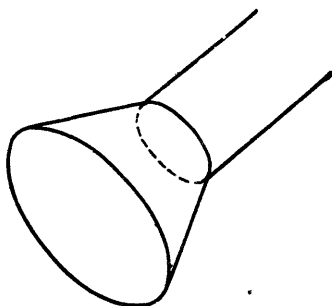


Fig. 6-60. Conical horn

for obtaining the long-distance communication at ultrashort waves) and with omnidirectional radiation in the horizontal plane. This can be achieved with the aid of biconical horns

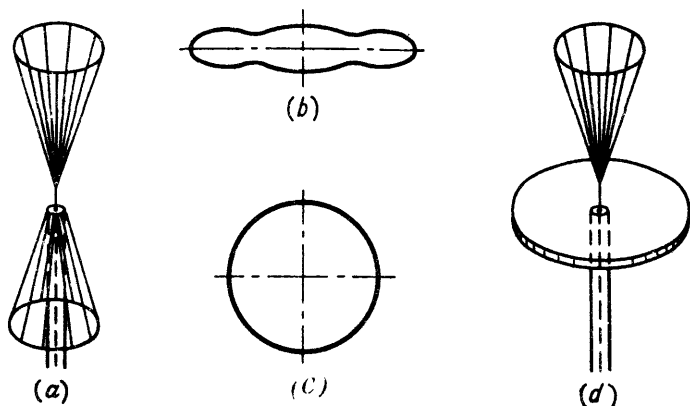


Fig. 6-61. Biconical antennas

(a) biconical horn; (b) radiation pattern in vertical plane; (c) radiation pattern in horizontal plane; (d) disccone antenna

where two conical horns (Fig. 6-61a) are fed from a coaxial cable. The greater the electrical length of the horn and the smaller its flare angle, the narrower the beam in the vertical

plane (Fig. 6-61b). In the horizontal plane (Fig. 6-61c) the antenna radiates uniformly in all directions. Instead of such a symmetrical antenna an asymmetrical antenna which has its lower half replaced by a conducting surface (disc) (Fig. 6-61d), thus producing the mirror image of the upper half, can be used. This type of antenna is known as a discone antenna.

Diffraction (slot) antennas can be of different types depending upon the shape of radiation pattern required. In design-

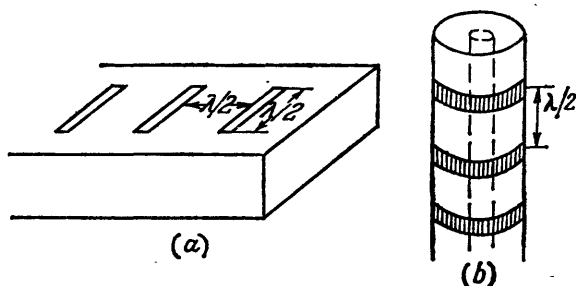


Fig. 6-62. Slot (diffraction) antennas

(a) inphased system of slots cut into waveguide; (b) inphased system of slots cut into coaxial feeder

ning such antennas the same principles are used as in the case of metal antennas. It is only essential to position the slots so that the magnetic lines of force will be oriented along them and the surface currents on the metal walls where the slots are cut intersect the slots at right angles. Fig. 6-62a illustrates a slot antenna presenting a system of inphased half-wavelength slots cut into the wide wall of a waveguide closed at one end. Maximum radiation of such a system is in the direction perpendicular to the wide wall. Fig. 6-62b shows an antenna consisting of a system of slots cut into the outer conductor of a coaxial cable. Such an antenna produces omnidirectional and directional radiation in the horizontal and vertical planes, respectively.

To excite a definite mode in a waveguide, an exciting system producing the waves of the required structure must be employed. For instance, to excite the dominant magnetic wave in a rectangular waveguide a linear vertical dipole (Fig. 6-63a) extended into the waveguide at the centre of its wide wall is generally used. The dipole is fed through a coa-

xial cable whose braid is connected to the waveguide wall. A movable reflecting wall is mounted behind the dipole at a distance approximately equal to a quarter-wavelength. Optimal excitation conditions can easily be provided by changing the position of this reflecting wall. Maximum intensity of the electric field is observed near the dipole and the field is directed along its axis which corresponds to the structure of the wave $H_{0,1}$ in the waveguide. To excite the dominant wave in a round waveguide, an exciting dipole is inserted into the waveguide at the centre of its end face (Fig. 6-63b).

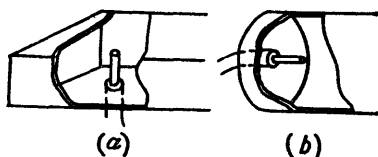


Fig. 6-63. Excitation of electromagnetic waves in waveguides

(a) excitation of dominant magnetic wave in rectangular waveguide; (b) excitation of dominant magnetic wave in circular waveguide

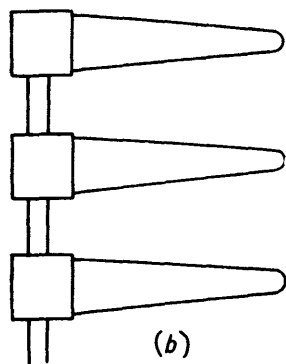
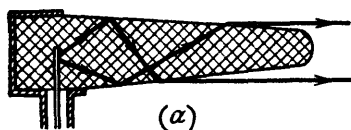


Fig. 6-64. Dielectric antennas
(a) radiator construction; (b) system of inphased dielectric radiators

Directional radiation can be produced with the aid of a *dielectric antenna* (rod) in which electromagnetic waves are excited by a small dipole (Fig. 6-64a). Behind the dipole is a reflecting wall. The rod has a variable cross section. Electromagnetic waves, reaching the rod — air interface, first undergo total internal reflection. As the rod narrows, the angle of incidence increases, and the waves start emerging from the rod at small angles to its axis. The shape of the rod is selected so that the waves can emerge from it at approximately equal angles to its axis, thus, producing directional radiation from the antenna. The directivity of the antenna increases with the increasing rod length. A beamwidth of about

30° can be obtained at a rod length of 5λ . To provide a narrower beam, a system consisting of several inphased rods is used (Fig. 6-64b).

At present *metal-plate lens antennas* are used to obtain very narrow beams. The operating principle of these antennas is based on the effect of increasing the phase velocity of electro-

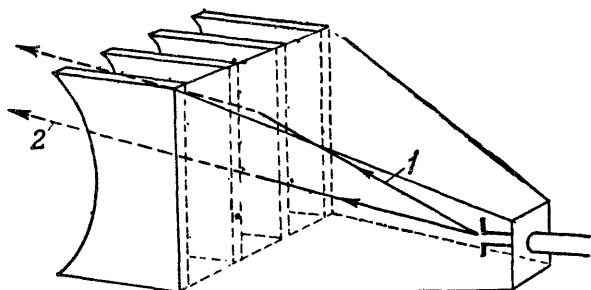


Fig. 6-65. Metal-plate lens antenna

magnetic waves propagating in the space between metal surfaces. Such a lens consists of several parallel metal plates of a specified shape. These plates are mounted just in front of a horn which directs the entire radiation from an exciting dipole (radiator) towards one side (Fig. 6-65). While passing through the lens the radiator's spherical wave is converted into a plane one. This is explained by the fact that side rays *1* travel a longer distance in the space between the metal plates than rays *2* propagating near the axis of the antenna. The plate shape is selected so that all the rays emerging at a given instant at different angles from the radiator leave the lens simultaneously. Therefore, the spherical wave front straightens out and the radiation pattern takes the form of a narrow beam. Lenses providing a beamwidth of a few minutes of arc are quite feasible. These radiating devices are particularly useful for radio-relay systems.

Chapter VII. RADIO WAVE PROPAGATION

7-1. PROPERTIES OF ATMOSPHERE AND GROUND AFFECTING RADIO WAVE PROPAGATION

In preceding chapters the reader has familiarized himself with the processes of radiation of radio waves, their reflection and propagation in the confined space of waveguides. We now direct our attention to the propagation of radio waves over great distances from a radiating device in the space through which this or other data (telephone or radiolocation signals, etc.) are carried by radio waves.

Let us disregard at first real conditions and imagine infinite "free" space, i.e., a medium characterized by a specific conductivity $\sigma=0$, magnetic permeability $\mu=1$ and a relative dielectric constant $\epsilon=1$. Assume that a device radiating radio waves linearly and uniformly in all directions is located in this space. Such a device should be called an isotropic radiator. Now assume that the radiation power of this device is equal to P_{Σ} .

A device designed for receiving data (information) being transmitted may be located at some distance r from the radiator. The question arises: will the reception of this data be assured?

To evaluate the influence of the radio waves upon the receiving device, it is necessary to know the intensity of electric field E produced by the waves at the reception point. As shown in the preceding chapter (Eq. 6-18), field intensity is proportional to the current through the radiator and inversely proportional to the distance between the radiator and the reception point. Considering also that the radiation power, in its turn, is proportional to the square of the current, the following relationship for an isotropic radiator can be obtained:

ned:

$$E \approx 5.5 \frac{\sqrt{P_{\Sigma}}}{r} \quad (7-1)$$

Here, E is the effective intensity of electric field expressed in volts per metre (V/m); P_{Σ} , the radiation power in watts (W); r , the distance in meters (m). Since the electric field gradually decreases with the increasing distance from the radiator, it may prove to be more convenient to express its intensity in microvolts per metre ($\mu\text{V/m}$).

Let us determine the field intensity produced in free space by a 10-kW isotropic radiator at a distance of 5000 km

$$E \approx 5.5 \frac{\sqrt{10^4}}{5 \times 10^3 \times 10^3} = 1.1 \times 10^{-4} \text{ V/m} = 110 \mu\text{V/m}$$

To evaluate this field intensity from the viewpoint of reception efficiency, the radio interference level at the input of the same receiver must also be taken into account. It should be noted that in practice a full-frequency range broadcasting is provided at received field intensities ranging from hundreds to thousands of microvolts per metre, whereas for special communication equipment field intensities of dozens and even several microvolts per metre prove to be sufficient. Special high-sensitive receiving devices employed in space research can receive even much lower signals.

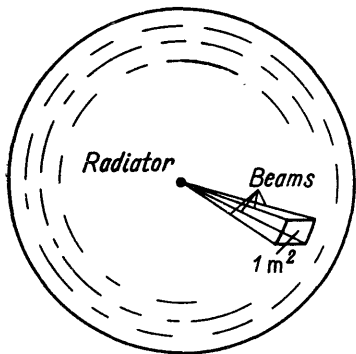


Fig. 7-1. Spherical wave front

Sometimes equation (7-1) is called the *ideal radio transmission formula*. This formula, however, is valid for Space-to-Space radio links only.

In free space the isotropic radiator produces spherical waves. These waves may be represented by spherical surfaces of equal field intensity for a given instant of time (or more precisely, identical amplitude and phase surfaces) whose centre coincides with the radiating point (Fig. 7-1). This is

quite evident since the waves propagate uniformly in all directions and the velocity of their propagation in free space is the same for any direction. The equal phase surface is referred to as a *wave front*.

The reason why field intensity decreases with the increasing distance is evident since radiator power is distributed over an increasing spherical surface as the wave moves away from the radiator and, thus, the amount of power per each square metre decreases (see Fig. 7-1). We may note here that a *small area of a spherical wave front at a great distance from the radiator may be considered plane*.

However, it is impossible and even impractical to produce isotropic radiation. As we know, actual antennas have directional radiation. It is characterized by the directivity factor D which indicates how many times the power flow in the direction of maximum radiation is greater than the average power flow (of course, at the same distance from the radiator). It will be recalled that the power flow (or the density of energy flow) is the power of an electromagnetic field per an area of one square metre, which is perpendicular to the direction of propagation (i.e., to a radio ray). The power flow is proportional to the square of electric field intensity (or to the square of intensity of the accompanying magnetic field).

For free space it can be easily shown that field intensity produced by an actual antenna in the direction of maximum radiation is given by

$$E \approx 5.5 \frac{\sqrt{P_{\Sigma D}}}{r} \quad (7-2)$$

Consequently, the antenna directivity effect is equivalent to increasing the radiation power D times, which is advantageous when the signal is received from the direction of maximum radiation.

If we wished to express the instantaneous magnitudes of the field of radio waves passing through the observation point, we would have to find the amplitude of the field intensity (i.e., $\sqrt{2} E$) and multiply this value by $\cos \omega \left(t - \frac{r}{v} \right)^*$.

* See Chapter 5, Eq. (5-4).

Here, $\omega = 2\pi f = 2\pi/T = 2\pi/v\lambda =$ radiator current angular frequency; $v \approx 3 \cdot 10^8$ (m/s) = velocity of propagation; λ (m) = wavelength. Hence,

$$E_{inst} \approx 7.75 \frac{\sqrt{P_{\Sigma} D}}{r} \cos \omega \left(t - \frac{r}{v} \right)$$

This equation describes the electric field of radio waves travelling in space.

In formulas (7-1) and (7-2) field intensity is independent of the wavelength. This is true only for free space although the radiation power itself P_{Σ} of any radiator depends essentially upon the wavelength.

If we now turn our attention from ideal conditions to real radio links installed on the earth, it will be found that the laws of propagation are intimately associated with the wavelength. First, radio waves propagate through the earth's atmosphere whose properties may differ significantly from those in free space. Second, radio waves propagate over the earth's surface which is convex and irregular and does not present either a perfect conductor or perfect dielectric. The earth's atmosphere and earth's surface greatly influence the laws of radio wave propagation and make the relationship between the field intensity and the wavelength at the reception points complicated. Therefore, field intensity at some reception point will be essentially different for waves of various wavelengths provided the antenna radiation power and directive gain are fixed.

The radio wave classification into frequency bands mentioned in the preceding chapters is primarily dictated by their propagation characteristics. These data serve as the basis in the assignment of available frequencies in the radio spectrum for specific purposes and applications.

The assignment of radio frequencies (radio waves) to certain frequency bands and their most typical applications are given in Table 7-1. Infrared and optical waves although included into the electromagnetic spectrum have much shorter wavelengths and do not fall into the category of radio waves.

It should be noted that metre, decimetre and centimetre waves fall so far into the category of "ultrashort waves" ("ultrahigh" frequencies). However, differences in their pro-

Table 7-1

Radio wave range	Wavelengths	Frequency range	Typical uses
Very long	10-100 km	30-3 kHz	Radio navigation, radio communication
Long	1-10 km	300-30 kHz	Radio communication, broadcasting
Medium	100-1000 m	3-0.3 MHz	Broadcasting, radio communication
Short	10-100 m	30-3 MHz	Radio communication, broadcasting
Metre	1-10 m	300-30 MHz	Radio communication, television
Decimetre	1-10 dm	3-0.3 GHz	Radiolocation, radio-relay communication, radio navigation, television
Centimetre	1-10 cm	30-3 GHz	Radiolocation, radio-relay service
Millimetre	1-10 mm	300-30 GHz	Special use
Submillimetre	Less than 1 mm	Above 300 GHz	Allocated for special use

propagation characteristics and particularly the design features of equipment operating in these three bands as well as first practical applications of millimetre and submillimetre waves make the term "ultrashort" waves go more and more out of use.

In the frequency bands of short, metre, decimetre and centimetre waves special sub-bands are allocated in the USSR for amateur radio communication and experimental work of radio amateurs. These are the following sub-bands: 3.5 to 3.65 MHz, 7 to 7.1 MHz, 14 to 14.35 MHz, 21 to 21.45 MHz, 28 to 29.7 MHz, 144 to 146 MHz, 430 to 440 MHz, 1215 to 1300 MHz, 5650 to 5670 MHz, 10 to 10.5 GHz, 21 to 22 GHz. A frequency of 27.12 MHz ($\pm 0.05\%$) is allocated for amateur controlled models. Thus, radio amateurs are given an opportunity to make their own contribution to radio engineering.

It is useful to remember that 1 gigahertz (GHz) = 10^3 megahertz (MHz) = 10^6 kilohertz (kHz) = 10^9 hertz (Hz).

Before describing the effect of the earth's surface and earth's atmosphere on the propagation of radio waves, we shall at first discuss the properties of the Earth and its atmosphere.

The Earth is an almost spherical body whose radius is about 6400 km. Consequently, the length of the earth's great circle is about 40,000 km. Therefore, the ground surface cannot be regarded as a flat one even for two points spaced several dozens of kilometres apart. If, for instance, points *A* and *B* are separated by 250 km (see Fig. 7-2), the height *h* will exceed 1 km and, therefore, radio waves cannot propagate in straight lines from point *A* to point *B*. Communication between points hidden from one another because of the earth's

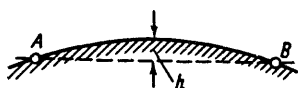


Fig. 7-2. Effect of Earth's curvature

curvature (as well as by surface irregularities such as mountains, hills, etc.) is possible on meeting one of the following conditions: either radio waves must propagate in curved lines (following around the earth's surface and

other obstacles), or transmitting and receiving antennas must be suspended sufficiently high to ensure "direct line-of-sight transmission", or, finally, intermediate stations capable of receiving and transmitting (relaying) radio signals should be installed between the sender and receiver.

The waves travelling along near the ground and following the earth's surface are called *ground* (or *surface*) waves. When propagating, ground waves encounter irregularities whose effect is determined by the height of the obstacle with respect to the wavelength. For instance, for long waves all surface irregularities (except mountains) are found to be less than the wavelength and, therefore, present no obstacle for the waves. In the decimetre and centimetre bands, however, even small ground vegetation or low seas are capable of preventing the propagation of radio waves. Particularly complicated for ground waves are the propagation conditions in cities where buildings cause multiple reflections.

Finally, the radio wave propagation is affected by the electrical properties of soils, especially in the vicinity of transmitting and receiving stations. If the ground were an ideal conductor (with specific conductivity $\sigma = \infty$), radio waves could not penetrate into the ground and, therefore, ground losses would be zero. The closest to such ideal properties are possessed by sea water over which propagation conditions for surface waves are most favourable. Wet ground is better

than dry ground which at long and particularly at short waves displays dielectric properties, thus causing displacement-current losses.

The earth's atmosphere is a gaseous envelope participating in the rotation of the Earth. It extends from the earth's surface up to about 20,000 km.

The lower portion of the atmosphere is called the *troposphere*. It contains 78% nitrogen (by volume), and 21% oxygen. Nitrogen and oxygen in the troposphere are constituted by conventional diatomic molecules (N_2 and O_2). In the lowest air layers atmospheric water vapour is also present.

Three main factors characterize the state and properties of the troposphere which affect radio waves propagation, these are: *air density*, *temperature* and *humidity*.

The density of the troposphere, i.e., the number of molecules per unit volume (per 1 cm^3) at a given altitude is proportional to the pressure of the above air layers. Of course, the highest pressure and highest density are observed near the earth's surface; they decrease with the increasing distance from the earth's surface. Near the earth's surface the pressure is approximately equal to one technical atmosphere, whereas at an altitude of 10 km it is about 4 times less.

The troposphere's temperature drops with the increasing altitude. This is explained by the fact, that air is heated not by sun rays; lower air layers acquire heat from the earth's surface which absorbs heat energy from the sun. The heated air rises and its place is taken by cold air. Consequently, this convection process causes a vertical displacement of air. On encountering horizontal flows of air (i.e., winds), the vertical air flows produce local *vortices*. This troposphere's property is known as *turbulence*.

The troposphere's temperature continues to drop up to some altitude and then remains relatively constant and equal to -60° to -70°C . It is precisely this temperature level which determines the upper boundary of the troposphere. This boundary lies at an altitude of about 11 km (or more precisely its altitude varies from 9 km for polar regions to 17 km for the tropics).

Humidity is the content of water vapour in the air measured on a percentage basis. Air humidity varies sharply with chan-

ging meteorological conditions and depends upon the underlying surface (dry land or water).

Thus, all the three troposphere's parameters (density, temperature and humidity) decrease with the increasing altitude following some average regularities. In some instances, however, certain deviations from these average regularities are observed due to varying meteorological conditions or turbulent phenomena. Local irregularities, for instance, may occur in the troposphere. All these factors affect to some degree the propagation of radio waves, resulting in the absorption of their energy and refraction.

Above the troposphere (about 90 km) air also consists of nitrogen (N_2) and oxygen (O_2) molecules which are mixed together by horizontal and vertical flows of gases. At altitudes exceeding 90 km nitrogen and oxygen molecules split into atoms and then form separate layers of these gases. Lighter monatomic nitrogen (N) becomes the basic gas at altitudes of 300 km and more. Needless to say that air density, i.e., the number of gas particles per unit volume at these altitudes is millions of times less than in the troposphere.

At altitudes of 60 km and more above the earth's surface gases are subject to *ionization* under the action of ultraviolet and X-rays, i.e., neutral molecules and atoms of oxygen or nitrogen split into positively charged ions and free electrons. This is very essential for radio waves propagation since the alternating electric field produced by radio waves will set the free electrons in motion and will interact with their field.

Ionization density, i.e., the number of free electrons per unit volume (per 1 cm^3), varies with altitude. One could think that the ionization maximum is close to the upper boundary of the atmosphere, but the gas density at these altitudes is so small that the number of electrons would be relatively small even if all the atoms were split. Maximum gas density is observed at the earth's surface which the ultraviolet and X-rays reach after being drastically weakened by passing through atmosphere layers. This means that the ionization maximum may be expected at some intermediate altitudes, which is actually true. Therefore, the atmosphere's regions lying at altitudes from 60 to 300-400 km above the earth's surface are known as the *lower ionosphere* while the *upper ionosphere* is at higher altitudes.

In the lower ionosphere several *regions* or *layers* exist, they are characterized by a certain degree of ionization. On one hand, this occurs due to different activity of ultraviolet and X-rays at various altitudes and, on the other hand, owing to the stratification of the upper atmosphere. Ionization is also caused by the impacts of material particles from the Sun on atmospheric gas molecules.

In previous years the ionosphere was studied with the aid of radio methods only. Ionospheric radio stations on the Earth radiated upward radio waves which after being reflected from the ionized regions were recorded automatically. Knowing the velocity v and measuring the time interval Δt taken by the transmitted wave to return, the altitude h (in metres) of the reflecting region can be determined

$$h = \frac{v\Delta t}{2} = 1.5 \times 10^8 \Delta t$$

In the daytime the radiated waves of different wavelengths were found to return from four altitudes and only from two at night. The concept, confirmed somewhat later, was advanced about the existence of several reflecting ionized layers. It was proposed that four ionization maximums called the *D*, *E*, *F*₁ and *F*₂ layers existed in the daytime; these layers are shown schematically in Fig. 7-3. The designation *F* was given to the last two layers. At night the ionizing action of sun rays ceases, free electrons recombine with ions and form neutral molecules. As a consequence, two layers *E* and *F* were considered to exist at night.

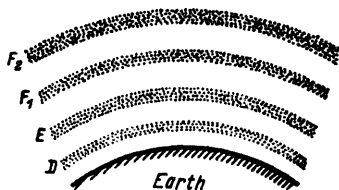


Fig. 7-3. Schematic representation of ionized layers (regions) of ionosphere

In the latter half of the fifties radio methods of ionosphere research have been supplemented by more accurate measurements made with the aid of geophysical rockets. The measurement results confirmed the previous concepts concerning the structure of the ionosphere. The electron density or, in other words, the concentration of electrons was found to increase by steps up to altitudes of 300-400 km and then to

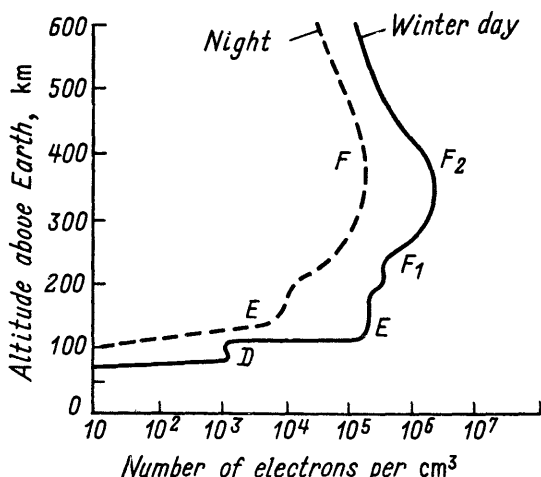


Fig. 7-4. Electron density as a function of altitude (approximate values)

decrease gradually. This is shown in Fig. 7-4 for day and night hours, respectively. It means that it is more accurate to say not about layers, but about various *degrees* or *regions of ionization* which alternately change.

A lower ionization region, the *D* layer, exists only in the daytime at altitudes of 60-90 km and is characterized by an electron density less than 10^3 per 1 cm³. At night the *D* layer disappears (see Fig. 7-4).

At altitudes of 100-140 km the daytime electron density rises up to 4×10^5 per 1 cm³. This is the *E* region. At night the ionization in the *E* region drops significantly, to 10^4 electrons per 1 cm³.

Both lower regions of the ionosphere (*D* and *E*) are characterized by irregular electron density. These irregularities arise, firstly, as a result of moving gas vortices, secondly, due to local ionization caused by material particles entering the earth's atmosphere from space. It is precisely at altitudes of 90-100 km that these material particles are heated as they collide with gas molecules and burn out, thus producing glow (meteors) and leaving behind ionized traces that then scatter for a few or dozens of seconds. Radio waves propagation may be affected both by basic ionization and its irregularities.

An increase in electron density up to 5×10^5 per 1 cm^3 is observed during the daytime at altitudes of 180-240 km, i.e., over the E region. This level is called the F_1 layer. It disappears completely at night. Finally, the region of maximum electron density called the F_2 layer is at altitudes of 230-400 km. On a winter day the electron density here is as great as 2×10^6 per 1 cm^3 . On a summer day owing to the rise of heated gases electron density in the F_2 region is not so high attaining only 2×10^5 per 1 cm^3 . At night the F_2 region has the electron density of about 2×10^5 per 1 cm^3 . It might be well to point out here that it is precisely the F_2 region which plays the major role in the long-distance propagation of short radio waves. As earlier the F_2 region may be called the F layer since electron density decreases both above and below it.

Apart from the time of day and season, electron density in the ionosphere (particularly in the F layer) depends upon solar activity, i.e., upon the intensity of ultraviolet radiation of the Sun. Solar activity can be judged by the number of sun spots which present vortical formations indicating an increase in temperature of the solar surface. From long-term observations it was found that solar activity has an eleven-year cycle (the last maximum of sun activity was observed in 1968 and the preceding one in 1957). A rise in solar activity increases the ionization levels in the earth's atmosphere followed by corresponding changes in radio wave propagation.

Directing our attention to the upper ionosphere, we briefly point out that electron density drops to 10^3 and 10^2 per 1 cm^3 at altitudes exceeding 1000 km and up to the boundary of the earth's atmosphere. These regions do not play so far such an important role for radio as does the lower ionosphere region. We have mentioned these regions here since they are closely associated with *ionospheric disturbances*.

As stated above, the earth's atmosphere is ionized, apart from ultraviolet and X-rays, by material particles (electrons, protons) generated during solar flares and by cosmic radiation. Just like a wire carrying a current, a moving charged particle interacts with the magnetic field of the Earth. In the result of this interaction the particle changes the direction of its motion and starts to rotate about the magnetic line of force (for more details see Chapter VIII). Following spiral trajec-

ries (Fig. 7-5a) with the decreasing radius of rotation as the field intensity increases, charged particles approach the Earth in its polar regions. A great number of particles moving over spirals about the magnetic lines of force form the so-called

radiation belt. A cross section of the radiation belt is schematically shown in Fig. 7-5b to give a simplified concept of its structure. Owing to the fact that particles from the Sun possess different energy, they are "detained" at different levels of magnetic field intensity. As observations showed, the radiation belt is made up of several (three) layers.

The charged particles invading (as a result of solar flares) the polar regions during collisions with gas molecules knock out photons, thus producing the phenomenon of Aurora. At the same time these particles cause a change in the normal density of ionization of the *F* region; therefore, this phenomenon is known as an *ionospheric or magnetic storm*. Magnetic storms cause changes in the general propagation regularities

based on the normal density of ionization in the *F* region, thus affecting radio communication (particularly in the polar regions) using this layer. Long-term observations (several years) of ionospheric disturbances provide sufficient material for their forecasting so that necessary precautions can be taken in advance.

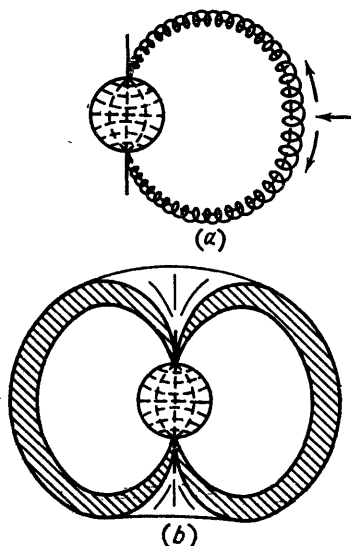


Fig. 7-5. Interaction between particles and Earth's magnetic field

(a) particles moving along helix trajectories; (b) cross section of radiation belt

7-2. RADIO WAVES PROPAGATING IN ATMOSPHERE. GENERAL REGULARITIES

Electromagnetic waves possess the property of bending around obstacles. The *bending* or *diffraction* of a wave is studied in physics with respect to the optical waves. When considering diffraction, one proceeds from Huygens principle which states: every point on an advancing wave front radiated by some primary source acts as a source that sends out a new (secondary) spherical wave. The longer the waves, the larger the obstacles they can bend around. Radio waves have longer wavelengths than light waves and, therefore, their diffraction is more pronounced.

As mentioned earlier, the earth's curvature presents an inevitable obstacle for signals transmitted over comparatively long distances. Only long radio waves are capable of following the earth's curvature. The researches of Academician V. A. Fock showed that even at very long wavelengths and at maximum possible radiation power the distance of diffraction propagation does not exceed 3000-4000 km. In practice, however, it is proved possible to provide communication at long waves even between the antipodes, i.e., over a distance of about 20,000 km. In order to explain such a possibility, it is necessary to consider the effect of the ionosphere which will be discussed below.

The earth's surface along which ground waves travel is characterized by local irregularities. As stated above, irregularities present obstacles for radio waves if their size appreciably exceeds the wavelength. Such irregularities cause a mirror reflection of radio waves. Smaller irregularities but still comparable with the wavelength ("rough" surface) result in *scattered reflection* of radio waves. A beam of incoming waves, which is oblique with respect to the average level of a rough surface is shown in Fig. 7-6. Scatter decreases the power flow which reaches a reception point. The less the height of irregularities and the smaller the angle of incidence of the beam, the less the scatter for the given wavelength.

Now assume that an asymmetrical grounded dipole (refer to Fig. 6-15) radiates radio waves over a smooth ground surface. If the surface layer were an ideal conductor ($\sigma = \infty$), the waves would be polarized strictly vertically (Fig. 7-7a).

A horizontal electric field at the surface of a perfect conductor is zero otherwise the field would produce an extremely high conductance current in the surface layer of the conductor. The lines of a vertical field terminate at the surface.

In actuality the surface layer is a semiconductor, i.e. it is characterized by a finite conductivity and some dielectric constant which differs from unity. The alternating electric field produced by radio waves penetrates the ground and expends energy there, heating the ground by conduction and displacement currents. Consequently, the beam, i.e., the direction of power flow,

becomes oblique (Fig. 7-7b): one portion of the energy is given off to the ground, and the other portion continues its motion along the earth's surface. However, the electric field intensity vector E always remains perpendicular to the beam. This means that the electric field vector also becomes inclined forward, towards the direction of radio wave motion.

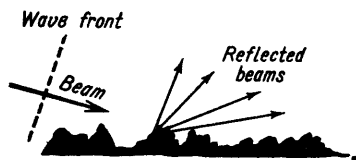


Fig. 7-6. Scattered reflection of radio waves by rough surface

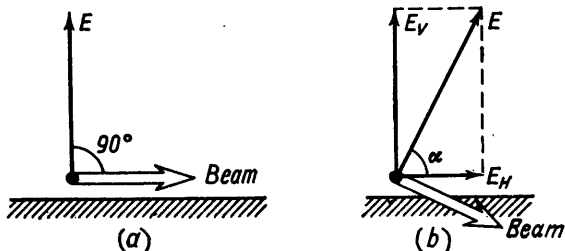


Fig. 7-7. Propagation of ground waves
(a) above ideal ground; (b) above actual ground

Vector E may be expanded into two components, *vertical* E_v and *horizontal* E_H . The greater the ground differs from an ideal conductor, the greater is the inclination of vector E and the lower the field intensity at the reception point as compared with the field intensity under ideal propagation conditions. Ground losses depend also upon the wavelength. For long waves propagating over the sea the conditions are

most close to the ideal ones. In general, *radio communication over great distances with the use of ground waves is feasible only at long wavelengths.*

One more peculiarity arises: a vertically polarized wave requires a vertical receiving antenna to produce an alternating current in it, whereas a field with oblique vector E having a horizontal component may also induce current in a horizontal receiving wire oriented along the direction of wave motion. A horizontal receiving wire is useful whenever a low suspended antenna is required.

On considering the influence of the earth's surface upon radio wave propagation, we return to the effect of the earth's atmosphere.

Assume that radio waves travel from the region of neutral gas to that of ionized gas. A transverse electric field (perpendicular to the beam) forces free electrons in the ionized medium to move. In its turn, the electromagnetic field produced by moving electrons interacts with that of radio waves and changes the direction of their propagation. In physics such a change in the direction of beams is called *refraction*. The transition from the neutral layer to the ionized one is characterized by a refractive index determined by the well-known relationship:

$$n = \frac{\sin \delta_1}{\sin \delta_2}$$

where δ_1 is the angle of incidence and δ_2 , the angle of refraction (Fig. 7-8) formed by the incident and refracted beams, respectively, with the perpendicular to the layer interface.

For radio waves the refractive index is proportional to the square root of the ratio of the dielectric constant of the second layer to that of the first. Taking into account that the dielectric constant for neutral gas equals approximately unity, we may write:

$$n = \frac{\sin \delta_1}{\sin \delta_2} \approx \sqrt{\epsilon_{i.g}} \quad (7-3)$$

where $\epsilon_{i.g}$ is the relative dielectric constant for ionized gas.

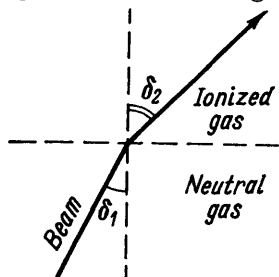


Fig. 7-8. Law of radio-wave refraction

If $\epsilon_{i.g} < 1$, then $\delta_2 > \delta_1$. This means that in an ionized layer radio waves bend more towards the Earth than in lower layers of the earth's atmosphere. This is illustrated in Fig. 7-8.

It should be recalled, however, that the dielectric constant of all solids, liquids and gases is greater than unity and only for vacuum it is equal to unity. Here, however, we assert that the dielectric constant of ionized gas may be less than unity, which means that ionized gas is a specific dielectric. How can we explain this physically?

When radio waves pass through an ionized medium, free electrons are set in motion by the field of these waves. Neglecting at the moment the possibility of collision between electrons and gas molecules or ions, one may assert the following: the higher the *ionization density* (i.e., the number of free electrons per unit volume) and the lower the *operating frequency* (i.e., the longer the distance travelled by an electron under the action of a half-wave), the stronger the action of an "electron" current in "suppressing" and ordinary displacement current. The resultant current decreases, which corresponds to a decrease in the dielectric constant of the ionosphere. Mathematically, this may be expressed as follows:

$$\epsilon_{i.g} \approx 1 - 81 \frac{N}{f^2} \quad (7-4)$$

Here, N is the number of electrons per cm^3 and f , the source frequency measured in kilohertz.

By combining formulas (7-4) and (7-3), we may write a refractive index expression for radio waves:

$$n = \frac{\sin \delta_1}{\sin \delta_2} = \sqrt{1 - 81 \frac{N}{f^2}} \quad (7-5)$$

This formula will be extensively used in analyzing the processes of wave propagation in the ionosphere. At present we note only the most important practical corollaries of this formula.

If ionization disappears due to recombination of electrons with ions ($N \rightarrow 0$), refraction ceases as well. Long waves (low frequencies f) bend more than short waves and for all ultrashort waves (very high frequencies f) refraction is absent even with the maximum possible degree of ionization being observed. Finally, at some value of the angle of incidence δ_1 (when

incident waves form a sufficiently low angle with the layer interface), the angle of refraction δ_2 would possibly become equal to 90° . This means that starting from this angle δ_1 the waves will not penetrate the ionized medium and will be completely reflected back to the Earth at lower angles of incidence.

Summing up what has been said above, we may assert that the longer the wavelength and the higher the ionization density of the layer, the greater the angle of refraction and the steeper the angle of incidence at which radio waves are reflected to the Earth.

For the case of vertical incidence ($\delta_1=0$), formula (7-5) will look as follows:

$$0 = \sqrt{1 - 81 \frac{N}{f^2}} \quad (7-6)$$

The value of frequency satisfying this condition is called the *critical frequency* of an ionized layer (f_c). It can be easily found from formula (7-6), i.e.,

$$f_c = \sqrt{81 N} \quad (7-7)$$

Consequently, instead of the number of electrons per unit volume the critical frequency f_c can be used to characterize an ionized layer.

For instance, let us find the critical frequencies of the F layer for day and night hours in winter (see Section 7-1). In the daytime

$$f_c \approx \sqrt{81 \times 2 \times 10^6} \approx 13,000 \text{ kHz} = 13 \text{ MHz}$$

At night

$$f_c \approx \sqrt{81 \times 2 \times 10^5} \approx 4000 \text{ kHz} = 4 \text{ MHz}$$

Thus, the critical frequency can change substantially during a day.

The real ionosphere has, of course, non-uniform ionization density as shown in Fig. 7-4. The density of ionization increases with the increasing altitude. On reaching the maximum, it then gradually decreases. One may visualize that ionization increases by steps (Fig. 7-9), so that $\epsilon_0 > \epsilon_1 > \epsilon_2 > \epsilon_3$. A radiated wave will undergo gradual refraction bending more and more towards the Earth until it finally reaches the

next ionization level (ϵ_3) with an angle of incidence at which it can be reflected back to the Earth. Then undergoing the inverse refraction process, the wave returns to the Earth at the same angle θ at which it has been radiated. Clear that the wave is reflected to the Earth before it has passed the

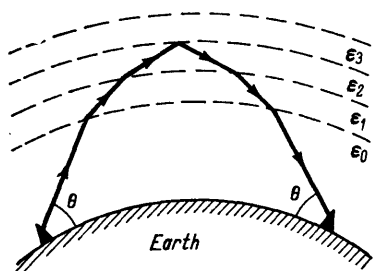


Fig. 7-9. Path of ray bending to Earth

maximum of electron density in a given region. If the density of ionization is not high enough to return the wave back to the Earth, it passes into space with its path somewhat bent in the ionosphere.

Figure 7-10 illustrates the most typical cases of the propagation of long (low radio frequency), short (high radio frequency) and ultrashort (ultra-high radio frequency) radio waves in the ionosphere. The reader may correlate these pictures with formula (7-5) and substantiate the following facts: roughly speaking, long waves are reflected from the lower boundary of the ionosphere; short waves may return back to the Earth provided the ionization density is high enough; and ultrashort waves

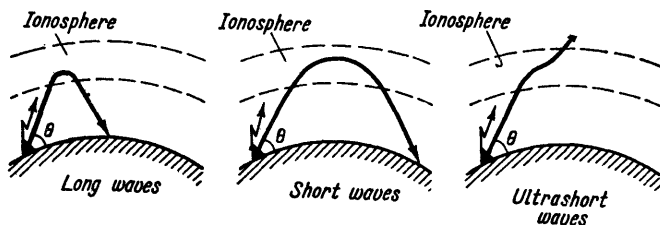


Fig. 7-10. Long, short and ultrashort waves propagating in ionosphere; typical cases

pass through the atmosphere into space. In the last case, the radio ray may undergo some displacement in the ionosphere but remains practically straight in space as well. Deviations from these general regularities will be specified in the following sections.

In conclusion it may be said that the oscillations produced by radio waves propagating in the ionosphere cause frequent collisions between electrons and gas molecules, thus some energy of radio waves is expended on heating the gas. This absorption of energy is a detrimental property of the ionosphere since it weakens the field intensity of radio waves. Characteristically, the greater the degree of ionization and the gas density, the greater the energy absorption, as the number of collisions between electrons and molecules per second rises with increasing gas density. Therefore, the *E* region (and even the *D* region) in which electron density N is less than in the *F* region but which is under a greater pressure absorbs the energy of radio waves more intensively. From this it follows, for instance, that for short waves the *F* layer is reflecting and the *E* and *D* layers are absorbing regions.

Absorption depends not only on the gas density but also on the operating frequency f . For wavelengths shorter than 300 m this relationship is very simple: the radio-wave attenuation decreases with increasing frequency. This may be explained physically. At a high frequency f an oscillating electron several times manages to acquire and reradiate energy before colliding with a gas molecule. The obtained and reradiated energy is returned back to the wave ("reactive" energy). This means that by the instant when the electron collides with the molecule, the energy stored by the electron for a short period of time will be small and decrease with the increasing frequency. Ultrashort waves, for instance, experience a relatively low attenuation in the ionosphere.

For wavelengths longer than 100 m propagating in the ionosphere an attenuation maximum exists which corresponds approximately to the medium-wave range. It is precisely in this wavelength range that the frequency of electromagnetic oscillations may prove to be equal to the number of collisions between the electrons and molecules per second. This results in a peculiar resonance phenomenon which correlates well with the maximum energy expended for "heating" the air. At much longer wavelengths attenuation is again somewhat reduced since electron transit time between two collisions is found to be less than the period of electromagnetic oscillations and, therefore, during the transit time the electron acquires only a small fraction of wave energy.

Thus, the ionosphere may behave as a medium which not only refracts radio waves but absorbs their energy, i.e., proves to be a semiconducting medium.

The wave which returns back to the Earth after being refracted in the ionosphere is called the *sky* or *space wave*.

7-3. LONG-WAVE PROPAGATION

It is well known, that very long and long waves ($\lambda > 1000$ m, $f < 3 \times 10^5$ Hz) display the maximum ability of following the earth's surface. At these waves communication is possible even between antipodes (i.e., over a distance of 20,000 km), which, however, cannot be explained only by diffraction since the presence of the ionosphere also favours their propagation.

Not only the *E* layer, but in the daytime also the *D* layer, has such an ionization density that long waves can reflect at any angle of incidence, including the vertical incidence. In fact, if we take even the highest frequency of the long-wave range ($f = 3 \times 10^2$ kHz) and substitute its value into formula (7-7), we obtain the electron density required for vertical reflection of the vertically incident wave: $N = f^2/81 =$

$$= 9 \times 10^4 / 81 \approx 10^3 \text{ electrons per } 1 \text{ cm}^3.$$

As we know, this figure corresponds to the daytime ionization density in the *D* region and is by two orders below that of the *E* layer. This means that the long waves are reflected from the lower layers of the ionosphere at any angle of incidence.

The surface of oceans and seas and even wet ground are found to be almost ideal conductors at the long waves, and, thus, they also reflect the waves at any angle of incidence. Therefore, the process of long waves propagation may be visualized as multiple reflections in the space between the lower boundary of the ionosphere and the surface of the Earth. This is identical to the propagation of waves in a gigantic

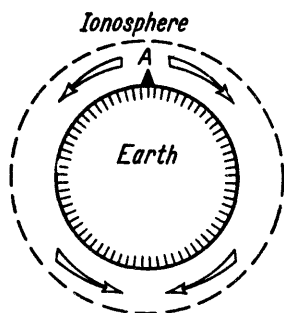


Fig. 7-11. Long-wave propagation and antipode effect

waveguide or even in a coaxial system (see, for instance, Fig. 5-27) where the ionosphere and the earth's surface (Fig. 7-11) are the walls of the waveguide. It is precisely this "guided" propagation that explains the reason why regular long-wave communication can be maintained over such great distances at which it would be impossible to obtain satisfactory radio-signal reception with the use of wave diffraction only.

The fact that in the long-wave range the field intensity at the reception point remains relatively constant during a day, a year and over an 11-year period, as well as the absence of ionospheric disturbances, is very important for communication and broadcasting at these waves. This fact is explained by the stability of the *E* region above which (in the region of ionospheric disturbances) long waves do not penetrate.

High attenuation of these waves, however, makes it necessary to use high power transmitters (with the output power equal to hundreds and even thousands of kilowatts). This is of particular importance because of the fact that long-wave antennas are generally omnidirectional.

A grave disadvantage for long-wave communication is a profound effect that atmospheric interference (static discharges) has on receivers. These discharges produce primarily long-wave noise. Finally, the bandwidth of long-wave communication systems is very narrow (due to the use of a low carrier frequency). As a consequence, wavelengths longer than 2000 m are not used in radiotelephony, though this fact has no direct concern with the mechanism of long-wave propagation.

7-4. MEDIUM-WAVE PROPAGATION

The medium-wave band (1000—100 m) has been used many years for broadcasting as well as for communication between merchant ships of many countries. The nature of propagation of these waves is no longer only of a waveguide type, since the ionization density in the *E* layer is not always high enough for their reflection and that in the *D* layer is even low. In fact, if we substitute the upper frequency limit of the medium-wave band ($f=3 \times 10^3$ kHz) into formula (7-7), we find the degree of ionization required for reflection of medium waves

at any angle of incidence, $N = f^2/81 = 9 \times 10^6/81 \approx 10^5$ electrons per cm^3 . This electron density is typical of the *E* layer only in the daytime. At night reflection can occur only at oblique incidence of the medium waves to the *E* layer boundary.

The following is remarkable: in the daytime the medium waves are highly attenuated in the *D* region through which they pass twice, when approaching the *E* layer and after reflection from it. At night the *D* region is absent so that no

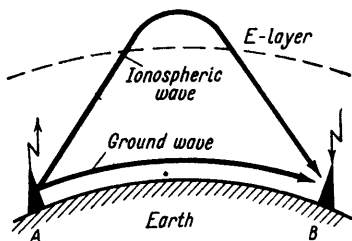


Fig. 7-12. Origin of fading

wave energy is absorbed in it. Therefore, in the daytime medium-wave communication can be accomplished practically through the ground waves only, whereas at night sky waves can increase significantly the transmission distance. This increase in coverage of medium-wave radio stations at night is very beneficial for

broadcasting. Medium-wave communication, however, must be maintained continuously (for 24 hours) which is feasible with the use of ground waves only.

At night the propagation of medium waves through the ionosphere is characterized by some specific features.

Fading (abrupt changes) in the intensity of the waves at the reception point is the first of these specific features. Assume that a transmitter at point *A* is sending out waves (Fig. 7-12) and a receiver is installed at point *B*. In the daytime only ground waves can reach point *B*, whereas at night it can also be reached by the waves reflected from the ionosphere. Therefore, the electric field at the reception point will be the result of interference (interaction) between the ground and sky waves. When these two waves are in phase, the resultant field strengthens and when they are out of phase it weakens (fades). However, the degree of ionization of the reflecting layer and, consequently, the depth of penetration of radio waves into it do not remain constant. They change randomly (or fluctuate) due to instability in the ionizing radiation from the Sun and the presence of air flows. This results in a change of path length of a sky wave and, consequently, in the phase shift between the ground and sky waves. The

shorter the wavelength, the more often are the changes in the phase relationship as the path length of the sky wave between the transmitter and the reception point fluctuates. Fading is possible in the zone which ground waves cannot reach and where interaction occurs between the sky waves approaching this zone by different paths. Such "long-distance" fading is frequently observed at short waves.

The second unavoidable consequence of ionospheric influence on the propagation of medium waves is the daily variations in the intensity of the waves at the reception point. At small distances from the transmitter where the field of ground waves is dominant, the received-signal strength hardly changes. At medium distances at which ground waves arrive highly attenuated, daytime reception may be poor, but at night reception improves, as the field of sky waves becomes dominant although the received signal fades. At great distances over which ground waves can hardly propagate, reception is possible only at night with sky waves.

The procedure for calculating the coverage of one or another medium-wave broadcast station (or, which is the same, the field intensity at the reception point) may be of interest for radio amateurs. It consists of the following steps.

First of all, one may agree to consider that the waves radiated by a grounded medium-wave antenna propagate in free "half space" limited at the bottom by an ideally conducting plane. These are the "ideal" conditions for ground wave propagation. In this case formula (7-2) will take the following form:

$$E \approx 7.75 \frac{\sqrt{P_{\Sigma} D}}{r} \quad (7-8)$$

Here, E is the effective value of electric field intensity (near the interface) at a distance r (m) from the transmitting antenna which produces radiation power P_{Σ} (W) and which has a directivity factor D in the horizontal plane. As we see, the effective value of the field intensity has increased $\sqrt{2}$ times, which corresponds to the doubling of radiation power. Physically this may be explained by the following. If the radiation power is maintained at the same level as was produced by the antenna in free space, then the entire power is directed to the half space and, therefore, the *power flow doubles*.

The field intensity considering the curvature of the Earth and actual energy losses in the surface layer will be less than the values calculated from formula (7-8). Field attenuation caused by bending around the Earth and the ground losses is accounted for by factor F which itself presents a proper fraction. Thus, we may write

$$E = 7.75 \frac{\sqrt{P_{\Sigma D}}}{r} F \quad (7-9)$$

The value of factor F depends upon the properties of the ground surface and wavelength. This factor is determined from reference tables based on calculations and experiments. Factor F has its maximum and minimum values for sea and dry sand surfaces, respectively.

7-5. SHORT-WAVE PROPAGATION

Generally speaking, the short waves (100-10 m) can propagate as ground and sky waves. It should be noted, however, that absorption of radio waves by the ground surface increases with increasing frequency, whereas in the ionosphere their absorption reduces with increasing frequency. It is precisely this fact that explains why ionospheric propagation is considered as the principal type of propagation used at short waves. Short waves propagate over comparatively small distances along the earth surface. Even with transmitter powers amounting dozens and even hundreds of watts the transmission distance achieved with ground waves does not exceed dozens of kilometres, which is particularly true for the upper half of the short-wave band (50-10 m).

Short-wave transmission with the aid of sky rays is an economical method for providing long-distance communication. The F region proves to be the basic reflecting layer for short waves under normal ionospheric conditions. The underlying E and D regions cause a severe attenuation of short waves. A "normal" passage of short waves is shown in Fig. 7-13a. Fig. 7-13b illustrates a possible increase in the transmission distance at short waves achieved owing to "two-hop" propagation, when the radiated wave is reflected twice from the ionosphere and once from the Earth. In this case, the transmission distance (or radio coverage) is determined by

the angle at which the short waves strike the ionosphere boundary (and are reflected from it)—the greater the angle of incidence, the greater the distance covered by one hop (also called the skip distance). The economy is achieved here owing to the fact that the operating wavelength can be chosen so that attenuation in the ionosphere will be insignificant for these short waves (much less than for medium waves). Therefore, at the points where reflected waves return to the Earth, their field intensity may prove to be sufficient for reception of signals even from comparatively low-power transmitters.

To explain how to select a suitable portion of the short-wave band, we consider Fig. 7-14 which shows the propagation of a short wave whose frequency is *higher than the critical frequency of the F layer*. We shall estimate the steepness of ray fall not by the angle of incidence δ_1 (see Fig. 7-8) but by the elevation angle θ formed between the wave ray and the tangent to the earth surface at the radiation point.

Waves sent out at steep angles ($\theta \rightarrow 90^\circ$) pass through the

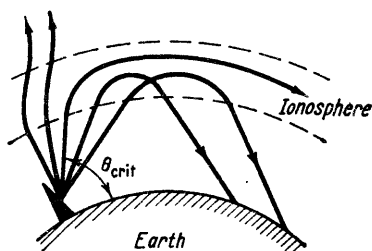


Fig. 7-14. Propagation of waves radiated at different angles; $f_{\text{wave}} > f_{\text{crit}}$

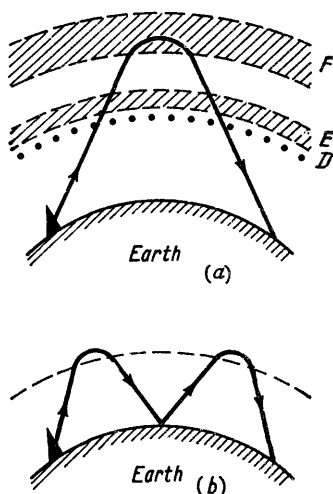


Fig. 7-13. Propagation of sky waves

(a) one-hop propagation; (b) two-hop propagation

ionosphere into space. Complete internal reflection occurs at a certain angle θ_{crit} (a critical angle for a given degree of layer ionization and a given operating frequency), and the ray is guided in the ionosphere parallel to the earth surface. At angles less than critical the rays return back to the Earth, with the skip distance being

the greater, the less is angle θ . The maximum skip distance of about 4000 km is generally obtained when waves are sent out in the direction tangent to the earth surface.

Thus, the desired transmission distance or coverage determines the *angle θ at which an antenna must radiate a maximum energy*. Knowing the height of the reflecting layer, it is easy to determine this angle from simple geometric considerations. To obtain sufficient sky wave field intensity at the desired reception point, the following two conditions should be met. First, the operating frequency (wavelength) should be chosen somewhat *lower than the critical frequency at which the waves are still reflected from the F layer at the required angle of elevation*. Second, *the attenuation of the waves of this frequency should not be too high during their two-way (up and down) propagation through the E and D regions* (it should be noted that attenuation is reduced with the increasing frequency).

Consequently, the choice of frequencies for short-wave ionospheric communication is limited to some frequency band. The upper tight limit is imposed by the critical frequency which is determined by the possibility of reflection, and the lower loose limit, by the lowest frequency which is determined by allowable attenuation. Both these limit frequencies depend on the time of a day (i.e., on the degree of *F*-region ionization) and on the ray path (i.e., on the elevation angle θ).

The "normal" picture of short-wave propagation described above helps one to explain the difference between "daytime" and "night" waves.

In the daytime the shortest waves of this band (approximately from 10 to 25 m) are used for long-distance communication, as they are capable of being reflected from the *F* layer even at low elevation angles. Although much longer waves can reflect from this layer even with greater success, high daytime ionization in the *E* and *D* regions results in extremely high attenuation, thus, an additional increase in transmitter power is required.

The lower portion of the short-wave band (approximately from 35 to 100 m) is used for long-distance communication at night since at this time of a day the ionization density of the *F* layer decreases and the shorter waves cannot reflect from it even at low elevation angles. At the same time attenuation in lower layers of the ionosphere is not so high at night since

the *D* region is absent and ionization of the *E* region greatly decreases.

The intermediate portion of the short-wave band (approximately from 25 to 35 m) is used successfully for communication at dawn and sunset. The boundaries for these three portions of the short-wave band cannot be specified accurately, since they depend upon the season (summer, winter) and the phase of the solar activity cycle.

The above discussion of "normal" short-wave propagation convinces us that power economy in providing long-distance communication at short waves is obtained at the expense of some operational complications, namely, the necessity of changing the operating frequency several times a day, and, what is more important, at the specified time. However, the "normal" conditions for short-wave propagation in the ionosphere are disturbed by a number of additional unwanted phenomena inherent in the short-wave communication.

As indicated earlier, the main kind of interference in the long-wave band is atmospherics or static. At the short waves, however, the most severe noise is produced by *interfering radio stations*, since the same frequencies suitable for long-distance communication during certain periods of a day are assigned to a number of radio stations whose signals propagate over great distances producing interstation interference. Needless to say that the required reduction of the noise from interfering (common-channel) transmitters presents certain problems.

Short-wave communication is also greatly affected by fading which is more severe and frequent at short waves than at medium waves. As before, this fading is caused by the *interference of two or more waves arriving at the reception point by different paths* with changing phase shifts. The time intervals between two minimums of the resultant field intensity may range from fractions of a second to several seconds and sometimes to dozens of seconds, whereas peak-to-peak field intensity (from minimum to maximum) may vary dozens and even hundreds of times.

As distinct from fading at medium waves caused most frequently by interference between ground and sky rays, at short waves fading is generally the result of interaction between two or more space rays which have traveled different

paths in the ionosphere. For instance, Fig. 7-13 illustrates reception of two waves one of which arrives at the reception point by one hop and the other, by two hops. Interference may also occur between the beams of waves radiated by a transmitter at different angles, which mutually overlap at the reception point. The techniques developed for reduction of fading will be discussed below. Needless to say that the need for reduction of fading presents additional

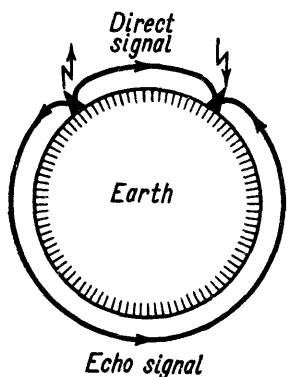


Fig. 7-15. Backward round-trip echo (separate reflections not shown)

problems in providing the ionospheric short-wave communication which is itself sufficiently complicated.

Radio-echo may produce additional interference in long-distance short-wave reception. Since attenuation of short waves is negligible, radio signals may travel round the Earth through multiple reflections so that a single signal may be received twice. Such a "forward round-trip echo" is observed 0.13 s after reception of the primary signal. In radiotelegraphy echoes may cause text distortions (particularly in case of high-speed operation); in facsimile reception echoes may cause ima-

ge distortions; and in radiotelephony, long-period reverberation.

A "backward round-trip echo", i.e., the signal which has traveled round the Earth in the direction reverse to the direct signal (see Fig. 7-15), may also be received. This echo-signal, however, can be eliminated by using the directional transmitting and receiving antennas.

Short-wave communication in polar regions (roughly speaking within 2000-3000 km from the poles) is frequently disrupted by ionospheric disturbances occurring primarily in the *F* layer. These cause an irregular decrease in ionization of the layer and may lead even to its complete disappearance for several hours. Additional layers may also appear or the height of the *F* layer may change. All these factors sharply weaken the field intensity of a signal rated for "normal" passage through the ionosphere.

To ensure reliable short-wave communication in the polar regions, it is advisable to design transmitters with a large power margin and to use highly directional antennas. Another method of providing the reliable short-wave communication is known as *diversity reception* in which two receiving antennas (and two receivers) spaced some distance apart are used. It is self-evident that long-wave and ultrashort-wave systems (if feasible) can provide more reliable communication in the polar regions than short-wave systems.

Short-wave communication conditions depend significantly on the phase of the 11-year solar activity cycle. The point is that in the daytime the *F*-layer critical frequency during periods of high solar activity is much greater (sometimes almost twice as great) than in the periods of low solar activity. *Approximate values of the F-layer critical frequencies* for middle geographical latitudes are given in Table 7-2.

Table 7-2

Phase of solar activity	Season	Time of a day	Layer critical frequency, MHz
Maximum	Winter	Day	14
		Night	2
	Summer	Day	7
		Night	4
Minimum	Winter	Day	8
		Night	2
	Summer	Day	6
		Night	4

Higher operating frequencies can be and should be used during periods of high solar activity, since attenuation in the ionosphere lower layers at these frequencies is lower, which is advantageous for long-distance communication. During periods of high solar activity, however, some *irregular phenomena may occur in the ionosphere* which may disrupt communication.

When organizing short-wave communication, one should take into account that a *zone of silence* (skip zone) may be ob-

served between the reception zones of ground and sky waves. Fig. 7-16 illustrates this phenomenon for a transmitter whose antenna produces omnidirectional radiation in the horizontal plane. The zone of silence is a ring-shaped area separating the two reception zones. The outer and inner boundaries of this area are determined by the minimum field intensity which still provides reception in the presence of noise background.

The existence of the silent zone can be explained rather

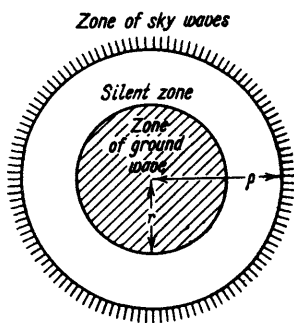


Fig. 7-16. Inner (r) and outer (ρ) radii of zone of silence (top view)

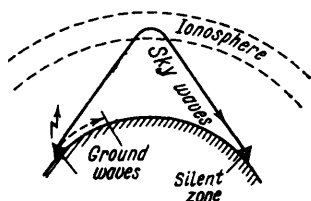


Fig. 7-17. Zone of silence

simply: ground waves (Fig. 7-17) cannot already reach this zone as they are fully absorbed by the ground, and the sky waves cannot reach this zone either as they are reflected only at the elevation angles which are not less than the critical one (see Fig. 7-14) and return back to the Earth beyond the silent zone. The inner radius r of the silent zone depends, of course, upon the radiation power of ground waves but it very seldom exceeds a hundred kilometres. At the same time the outer radius ρ , when a short-wave radio station transmits signals over great distances, i.e., when signals are radiated at low elevation angles, may exceed a thousand kilometres. For a given wavelength and a given transmitting antenna no radio signals can be received within the zone of silence even if the transmitter power is increased.

Thus, very interesting possibilities offered for communication by short waves can be realized only by considering thoroughly their properties and specific features.

For successful organization of long-distance short-wave communication one should know the regularities of short-

wave propagation and have at hand the required experimental data on the ionosphere. First of all the wavelength and the types of transmitting and receiving antennas are chosen. Assume a given radio path between points *A* and *C* (Fig. 7-18*a*). Suppose further that the height *BD* of the reflecting layer and the critical frequency of the region are known from the results of vertical ionosphere sounding made somewhere at the centre

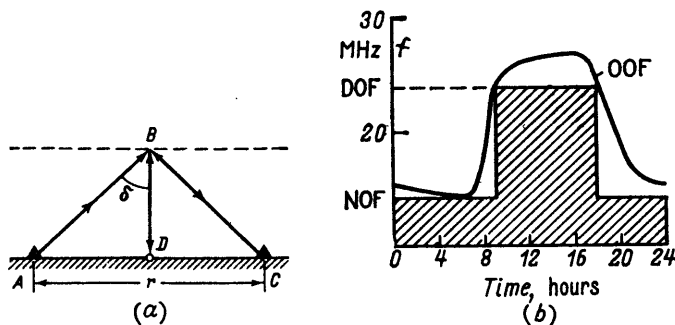


Fig. 7-18. Selection of operating frequencies for short-wave communication

(*a*) geometrical plotting; (*b*) daily variations in optimum operating frequency; *DOF* — day operating frequency; *NOF* — night operating frequency; *OOF* — optimum operating frequency

of this path for a certain period of a day. Here, it is more convenient to think of the wave as being reflected rather than gradually bent to the Earth. Then the maximum allowable frequency to be used for the given path during the required time of a day can easily be found from the geometrical relationships shown in Fig. 7-18*a*:

$$f_{m.a} = \frac{f_c}{\cos \delta} \quad (7-10)$$

where δ is the angle of incidence of the ray on the reflecting surface.

Assume, for instance, that distance $r=800$ km, height $BD=400$ km, critical frequency $f_c=7$ MHz. Then we obtain that $\delta=45^\circ$ and from formula (7-10)

$$f_{m.a} = \frac{7}{0.707} \approx 9.9 \text{ MHz}$$

This frequency corresponds to a wavelength of about 30 m. In practice, however, a somewhat lower operating frequency rather than the maximum allowable one is generally used. The optimum operating frequency (OOF) is generally 15% less than the maximum one. In the given example the optimum operating frequency was chosen to be equal to approximately 8.42 MHz. Knowing the angle of incidence δ , we may determine the desired elevation angles θ for orientation of the main beams (lobes) of the transmitting and receiving antennas and, thus, select the correct design. In our example triangle ABD is an isosceles one, so that the angles of elevation of the antenna main beam must be 45° for both antennas. Such a simplified construction, not taking into account the earth's curvature, may be tolerated, of course, for relatively short radio paths (hundreds of kilometres).

On selecting the operating frequency (i.e., the wavelength) and the types of antennas, one should determine the transmitter power (or more accurately the radiation power). Such a calculation can be accomplished on the basis of a method developed by A.N. Kazantzev.

All the above said about the selection of frequency, antenna patterns and radiation power applies to a certain period of a day and to a certain season. The daily variations in the critical frequency, however, cause changes in the maximum allowable frequency and, consequently, in the optimum frequency to be used. For instance, in Fig. 7-18*b* daily variations in the optimum operating frequency taken in autumn are shown for a radio path of about 1500 km. The upper part (higher frequencies) of the shaded area indicates the values available for selection of the day operating frequency (DOF) and the lower part (lower frequencies), for selection of the night operating frequency (NOF), both approaching as near as possible the optimum frequencies.

These are the most elementary concepts concerning the organization of short-wave communication using sky waves. Practical measurement methods are also used in addition to the calculation methods for determining operating frequencies assigned to preset radio paths. The former methods are based on the results gained from vertical sounding of the ionosphere. Direct measurements are based on the *backscattering effect* of sky waves as they fall on the earth surface.

Surface irregularities, even if they are small in size, produce scattered reflection (Fig. 7-19). If at transmission point *A* a receiver is also installed and if this receiver detects waves reflected from the ground surface at point *B*, one may say that direct signals also travel along path *AB*. The fact that scattered reflection occurs precisely at the desired point *B* is established by measuring the time required for the direct and reflected signals to travel between the two points. Oblique sounding of radio paths with the aid of special equipment is a practical method used rather extensively.

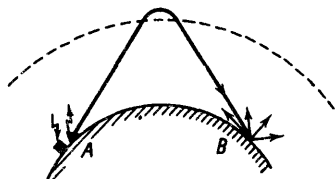


Fig. 7-19. Oblique sounding

7-6. ULTRASHORT-WAVE PROPAGATION

Ultrashort waves offer considerable scope for application of radio communication and broadcasting. The engineering fields such as television, radiolocation, short-range radio navigation and many others are proved feasible only with ultrashort waves because of the wide band of radio systems and highly directional antennas.

The application of ultrashort waves has been confined for many years to the requirement for line-of-sight (or direct ray) transmission between the transmitting and receiving antennas, which is a consequence of essentially straight-line propagation of these waves. In fact, diffraction around large obstacles is almost uncharacteristic of ultrashort waves and, therefore, they are hardly able to bend around the earth's curvature. On the other hand, at ultrahigh frequencies the sky wave is no longer reflected back to the Earth from the ionosphere because of a low degree of its ionization (except for periods of high solar activity when reflection is possible for the lower portion of the metre band ranging approximately from 7 to 10 m). It was precisely for this purpose that ultrashort waves were initially used for short-range ground communication, local radio and television broadcasting as well as for communication between aircraft.

However, science has found ways of solving the contradiction between the requirement for a wide bandwidth and the limited range of line-of-sight transmission. A *radio relay system* was developed which uses a number of intermediate stations (radio repeaters) along the radio path at line-of-sight intervals between adjacent antennas (Fig. 7-20). It is true, however, that radio relay systems prove to be more similar to wire communication than to the radio long-range communication both by performance and cost.

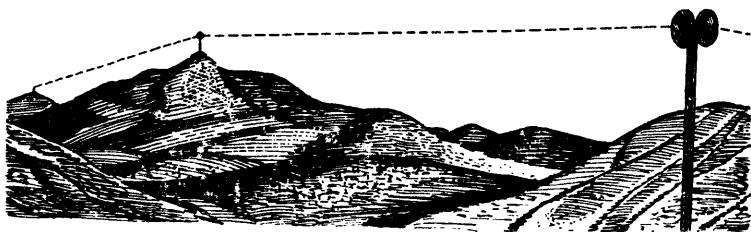


Fig. 7-20. Radio relay system using line-of-sight transmission

An intermediate station of a radio relay system receives radio signals incoming from one direction, and then reradiates (relays) these signals automatically in another direction at another frequency. By this means, radio signals sent out by two terminal stations are transmitted in both directions "through a chain" of intermediate stations. Moreover, one and the same antenna can be used for transmitting and receiving signals in one direction provided these signals have different polarization (horizontal or vertical). It is precisely this fact that explains why each tower (mast) shown in Fig. 7-20 has two (not four) parabolic antennas.

Radio relay systems employ so-called *multiplex methods*, that permit the simultaneous operation of *two or more telephone channels*, i.e., several pairs of subscribers may talk at the same time. Large systems have up to one thousand and even more channels. Wide band ultrashort-wave equipment makes it possible for radio relay systems to transmit not only telephone signals but also television programs between different cities and countries.

Ultrashort waves propagating through the lowest atmospheric layer undergo absorption which weakens their field

intensity as compared with the field intensity in free space. Water drops in the form of rain or fog may be one of the reasons for the wave absorption. This absorption, however, becomes significant only at wavelengths shorter than 3 cm.

The process of wave absorption by water drops can be explained physically as a result of inducing displacement currents in them, with current density being proportional to frequency. In addition, the boundary of a rain zone or cloud, as a medium having other parameters (for water $\epsilon_r=80$), cau-

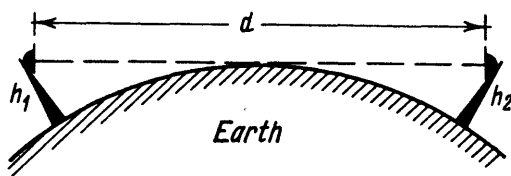


Fig. 7-21. Maximum line-of-sight distance between antennas

ses reflection and scattering of very short waves. Therefore, a communication system operating at a wavelength of 2 cm becomes unreliable during atmospheric precipitation. Appreciable absorption of energy in water vapour molecules and then in oxygen molecules is observed at millimetre waves.

For the reasons given above, waves not shorter than 3 cm are used for radio transmission along the ground surface. Television, for instance, uses metre wavelengths and in the recent years, decimetre wavelengths. Radio relay systems operate primarily at wavelengths from 3 to 10 cm and also at decimetre wavelengths. Communication between mobile vehicles on the ground (cars and rail transport) is accomplished basically at metre wavelengths.

Straight-line propagation of ultrashort waves requires that both transmitting and receiving antennas be located as high as possible. In stationary installations special supports (either towers or masts) are used for mounting ultrashort-wave antennas. Suppose that the earth surface is ideally spherical and smooth (Fig. 7-21) and that the height of the first and second antennas is h_1 and h_2 (in metres), respectively, then the maximum line-of-sight distance d (in kilometres) may be determined from

$$d = 3.57 (\sqrt{h_1} + \sqrt{h_2})$$

For example, if $h_1 = h_2 = 25$ m, the line-of-sight distance will not exceed 35.7 km. However, the wave sent out by the transmitting antenna to the receiving one propagates within some zone which includes a dashed line shown in Fig. 7-21 (in physics this zone is called the first Fresnel zone). This means that "clearance" of at least some metres is desirable between

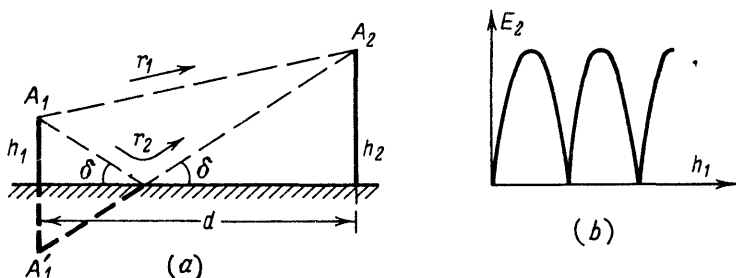


Fig. 7-22. Interaction between direct and reflected waves at reception point

(a) geometrical plotting; (b) resultant field as a function of transmitting antenna height

the highest point of a possible obstacle and the straight line passing through the antennas.

The antennas of the most important radio relay systems are mounted on towers up to 100 m high. TV transmitting antennas designed to serve the greatest possible area are raised much higher. For instance, the height of the Ostankinskaya Television Tower in Moscow is 525 m.

In the case of line-of-sight transmission between ultrashort-wave stations, the wave reflected from the earth surface may be present at the receiver input in addition to the direct wave. Fig. 7-22a illustrates a transmitting antenna A_1 of height h_1 and a receiving antenna A_2 of height h_2 . A direct wave ray is a straight line passing through the antennas. The reflected-wave ray can easily be found by connecting the transmitting-antenna image A_1' with the receiving antenna A_2 with a straight line. This gives the reflection point on the ground surface at which the angles of incidence and reflection (δ) are equal to each other.

The length of the direct ray r_1 is less than that of the reflected ray r_2 . Therefore, waves arrive at the point A_2 by two dif-

ferent ways, generally speaking, with different phases. The phase difference depends upon the path difference ($r_2 - r_1$) and the wavelength. For a fixed wavelength the path difference varies with the heights of antennas and the distance d between them. If any of these three parameters is continuously changed, the phase difference will vary periodically from 0 to 180° and then down to zero, etc. Consequently, the field intensity E_2 will also vary periodically as shown in Fig. 7-22b.

From this property discovered by B.A. Vvedensky a useful consequence follows, namely, if in a properly designed radio link the ray r_1 must ensure, according to the calculations, reliable communication but no communication is actually possible or proves to be unreliable, then good results can be obtained by varying slightly one of the above parameters (h_1 , h_2 , d or λ). The effect of interference between the direct and reflected rays may also be observed at a distance d at which the earth's curvature cannot be disregarded. On the other hand, an appreciable portion of energy of the ray r_2 may be absorbed or scattered during reflection and, therefore, the resultant field will be nonzero even with a 180° phase shift.

This concludes our short summary on ultrashort-wave propagation through the lowest atmospheric layer under conditions of line-of-sight transmission between corresponding antennas. With meter waves, which are more capable of bending around small obstacles on the earth surface than decimetre waves, communication is possible even beyond line-of-sight distances. Such communication frequently called "ground" communication is extensively used by radio amateurs. At meter waves, the ground-communication range is about several kilometers and seldom reaches dozens of kilometers. Most complicated conditions for meter-wave communication are within *large cities* where line-of-sight transmission between mobile radio stations is difficult to provide. Therefore, a "central" radio station with an antenna mounted on a high building is frequently used for relaying signals between the mobile stations.

Ultrashort waves offer very interesting possibilities for transmission of radio waves *beyond the horizon*. This effect was discovered in the latter half of the present century. It is explained by specific atmospheric properties causing ultrashort waves to propagate in different paths.

In some instances the index of refraction for air decreases with altitude much quicker than under normal atmospheric conditions. This is often observed over a water surface since air humidity in the lower troposphere layers over water is

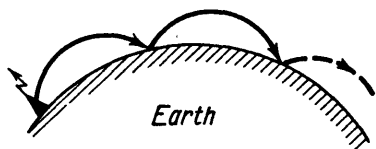


Fig. 7-23. Ultrashort waves propagating in tropospheric waveguide

high and decreases sharply with the increasing altitude. Under these conditions radio waves are bent sufficiently to be returned to the earth, be reflected from it, undergo refraction in the troposphere once again, and return to the earth at a doubled

distance, etc. (Fig. 7-23). Such cases of long-range propagation of ultrashort waves caused by tropospheric superrefraction are known as propagation in a *tropospheric waveguide*.

Long-distance reception of television programs due to tropospheric superrefraction may be of great interest for televiewers.

Tropospheric scatter, i.e., the scattering of radio waves caused by *local irregularities of the troposphere* is another possibi-

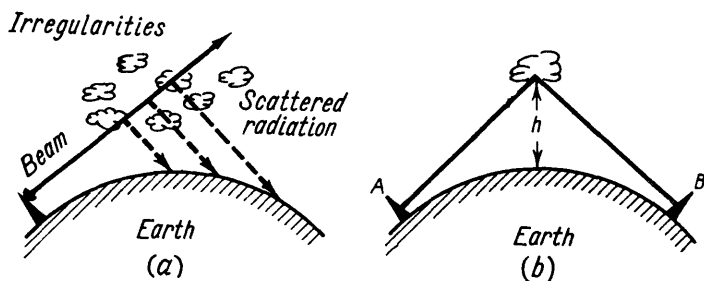


Fig. 7-24. Ultrashort-wave communication using scattered reflection
(a) scattering in irregularities; (b) range of tropospheric communication

lity for long-distance communication at ultrashort waves which proved to be regular and comparatively reliable. As we know, such irregularities constantly appear and disappear due to turbulent air flows at altitudes of 3-5 km and over. From the electrical viewpoint, these irregularities are characterized by a refractive index differing from the normal one.

While passing through the tropospheric irregularities, the ultrashort waves undergo partial reflection (Fig. 7-24a). These scattered waves are directed primarily forward and down so that some part of the radio wave energy reaches the earth at distances of hundreds of kilometers. The transmission distance in this case can be calculated from the known height h of the tropospheric irregularity by constructing two tangents to the earth surface (Fig. 7-24b). For instance, if $h \approx 3$ km, then the distance (the length of arc) AB can be calculated geometrically, which is about 400 km. If radio stations are located at points A and B we obtain an ultrashort-wave communication link using scattered reflection in the troposphere. Such links are generally designed to provide transmission distances of up to 200-300 km but in some instances even much greater distances (up to several thousands of kilometers).

It is much more difficult, however, to build the tropospheric communication link than a short-wave system providing the same transmission distance. Owing to the scattered reflection of waves from tropospheric irregularities, the field intensity decreases more sharply with distance than in the case of line-of-sight transmission. Therefore, high power transmitters (from 1 to 50 kW), highly directional antennas and high-sensitive receivers are required. It was only after creating such equipment that the phenomenon of tropospheric scatter was discovered.

The frequency range suitable for tropospheric communication includes meter, decimeter and certain centimeter waves (which are characterized by low attenuation in the atmosphere). The existing (known from technical literature) tropospheric communication lines operate primarily in the decimeter band.

Tropospheric communication employing scattered reflection is very useful because of the following properties. First, as distinct from long-wave and short-wave communication, a very wide band of frequencies can be transmitted with ultrashort waves, i.e., a great number of channels can be provided. Second, the ultrashort-wave operation requires no wavelength changes during the time of a day and, thus, is more advantageous than the short-wave one. In the tropospheric-communication link, however, the received signal intensity varies with time. A more stable signal level can be assured by

using two receivers with their antennas spaced some distance apart as in the case of short-wave reception.

Since the late fifties the tropospheric-communication links have been extensively used, competing, over almost inaccessible places, with cable lines whose running under these conditions would be more expensive. In radio-relay systems tropospheric stations may be spaced not at intervals of 40-50 km but at a distance of 150-250 km or more from one another which is also important in rather inaccessible areas.

Theory and design methods for tropospheric-communication links have been discussed in the works of B.A. Vvedensky, M.A. Kolosov, A.I. Kalinin, Yu.I. Davidenko and other scientists.

Scattered reflection of ultrashort waves and, particularly, meter waves is also observed in the lower layers of the ionosphere.

Ultrashort-wave ionospheric stations are used for communication between continents and remote islands in the absence of submarine cables or as a duplicating means. For this purpose, high-power transmitters (dozens of kilowatts) and inphased multidipole antennas are required. Among the advantages of these systems, as compared with short-wave communication systems having the same transmission distance, is the possibility of a 24-hour operation without changing the wavelength. During periods of ionospheric disturbances when normal short-wave communication based on the reflection from the *F* layer is disrupted, the ionospheric communication using meter waves is even improved since the irregularities present in the lower ionospheric layers become more pronounced.

7-7. ELECTROMAGNETIC WAVES IN OUTER SPACE

When dealing with a conventional ground radio system, the action of the transmitted signal on a receiving antenna was estimated by the electric field intensity produced by radio waves at the reception point (see formula 7-2). Although being convenient for radio broadcasting where signals can be received by antennas of any type, such an estimation cannot be applied for special radio systems. The point is that the signals used in such systems may have the form of separate (discrete) radio pulses; their action on the receiver is determined not only by the

field intensity but also by the pulse length and pulse repetition frequency. In this case, the action of the transmitted signals on the receiver is characterized more accurately by the power P_2 which the receiving antenna is able to extract from the radio waves.

In the general case, however, the ratio of the transmitted power P_1 to the power P_2 applied to the receiver input is used instead of the absolute value of P_2 . This ratio may be called *path attenuation*.

Directing our attention to the propagation of radio waves in *outer space* and assuming that losses due to collisions between the electrons and gas molecules are completely absent, we obtain the formula:

$$\frac{P_1}{P_2} = \left(\frac{4\pi r}{\lambda} \right)^2 \frac{1}{D_1 D_2} \quad (7-11)$$

Here, r is the distance from the transmitter to receiver, metres; λ is the wavelength, metres; and D_1 and D_2 are the directivity factors of the transmitting and receiving antennas, respectively.

From formula (7-11) it appears that longer waves are more advantageous since attenuation decreases with the increasing wavelength λ . This means that a given power P_2 can be provided with a lower power P_1 . However, we know that antenna directivity factors D_1 and D_2 can be increased by decreasing the wavelength. In most cases this proves to be the decisive factor in selecting the operating wavelength.

Let us calculate, for instance, the directivity of a transmitting and receiving antenna for a Space-to-Space radio link so that a received signal power of 10^{-11} W is ensured at a distance of 300,000 km with a wavelength of 3 cm and a radiation power of 1 kW. Assuming that $D_1 = D_2 = D$, from formula (7-11) we have

$$D = \frac{4\pi r}{\lambda} \sqrt{\frac{P_2}{P_1}} = \frac{12.56 \times 0.3 \times 10^6 \times 10^3}{3 \times 10^{-2}} \sqrt{\frac{10^{-11}}{10^3}} = 12,560$$

Such a high directivity can be obtained (with tolerable antenna sizes) only in antennas operating at very short wavelengths.

The above method for calculating the path attenuation is valid, strictly speaking, only for outer space (i.e., beyond the atmosphere).

At present radio engineers solve three major problems associated with radio-wave propagation in outer space. The first consists in serving ground listeners and viewers with programmes transmitted over great distances through artificial satellites by radio relaying. The second is to control

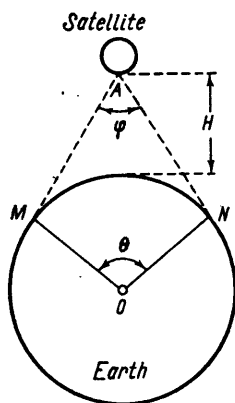


Fig. 7-25. Illumination area provided by artificial satellite at height H

spacecraft from ground stations and to provide communication between astronauts and the Earth. Finally, the third involves communication and mutual control between spacecraft beyond the earth's atmosphere without signals sent from ground stations. The solution of each problem has its own specific features depending to a great extent upon the radio-wave propagation conditions.

The usage of a satellite-retransmitter (such as the Soviet "Molniya" satellite) is illustrated in Fig. 7-25. Such a satellite carries aboard equipment which is identical with that of an intermediate station in a radio relay system. It includes two receivers and two transmitters as well as external antennas.

The purpose of the satellite is to receive signals from one ground station and automatically reradiate them to another. The maximum distance between the two stations equals the length of arc MN which corresponds to central angle θ . Points M and N are the tangency points of straight lines AM and AN drawn from the satellite to the Earth. The ground surface area which falls within solid angle φ is referred to as the "illumination area". In principle, communication may be maintained between any pair of stations located within the illumination area. In turn, the size of the illumination area is determined by satellite height H .

Only one receiver and one transmitter of the satellite-retransmitter operate if it is used for transmitting central television and radio broadcasting programmes to remote regions

of the USSR. Needless to say that communication and radio broadcasting through the satellite-retransmitter are possible only at that time of a day when the satellite is within the field of view of both ground points, i.e., when both points are located within the illumination area. To provide a "stationary" illumination area on the earth surface, the satellite-retransmitter must be launched to an altitude of about 36,000 km over the equator. In this case, its angular speed will be equal to that of the Earth, and the satellite will seem to be stationary to a ground observer. The problem of orbit selection is of great interest but it lies outside the framework of radio-wave propagation.

What wavelengths may be selected for retransmission of signals through a satellite? Taking into account that radio waves pass twice through the entire thickness of the atmosphere and undergo absorption in it, we have to abandon the waves shorter than 2-3 cm. At the same time, since ground receiving antennas are directed upward one should consider the action of interfering radiation from the Sun, cosmic bodies and interstellar gas. This influence is particularly severe in the meter-wave band, therefore, meter waves are also used very seldom for satellite communication. Decimeter waves and especially the lower portion of the centimeter-wave band (3-10 cm) are also used for this purpose.

The second problem is communication between the Earth and astronauts as well as flight control of spacecraft from ground stations. The presence of a man in a spacecraft imposes more stringent requirements upon the radio links than the requirements for unmanned space flights. The radio link must ensure voice communication (radiotelephone) between astronauts and ground control centres since it is precisely voice communication which makes it possible to report any situation, including emergency. At the same time, the radio link is used for transmission of control and command signals as well as telemetering data. Finally, the possibility of a spacecraft landing at any point on the Earth and the need for finding it with the aid of radio direction-finding equipment is also considered.

All these circumstances require the installation of both short-wave and ultrashort-wave radio equipment aboard the spacecraft. Short-wave signals from the spacecraft can be received by ground centres at great distances. The quality of their re-

ception, however, depends upon the ionospheric condition and it is often impaired by fading. Therefore, several centres positioned at different points on the ground and on the water areas generally receive radio signals from spacecraft.

Now we return to the third problem involving communication and mutual control between spacecraft without signals sent from ground stations. In this case, there are no limitations in the selection of wavelengths which were dictated by the properties of the earth's atmosphere. Millimeter, submillimeter and even optical waves in principle may be used for Space-to-Space communication and control links. Highly directional optical radiation produced by lasers is of particular interest. The choice of operating frequencies depends here upon complexity of the equipment and methods of antenna-beam orientation.

The actual radio control and radio communication links used during interplanetary flights indicate great future possibilities and the tremendous achievements in the practical realization of laws of electromagnetic wave propagation.

Chapter VIII. VACUUM AND SEMICONDUCTOR DEVICES

8-1. MODERN ELECTRONICS

The study of electron devices constitutes an essential part of the modern science and technology, which is called electronics. This subject deals with the properties and laws of motion of electrons as well as with conversion of the different forms of energy through the medium of electrons. On the basis of the laws of electron motion devices were created being widely used in different fields of science, technology and life, in various mechanisms and machines. Radio engineering is one of the fields of applied electronics which invented various devices such as vacuum devices, semiconductor devices, solid-state devices, lasers and masers.

Electron vacuum devices are based on the motion of free electrons and ions in vacuum or rarefied gases under the action of electric and magnetic fields. These devices are used in the majority of radio electronic equipment. Their role is of particular importance in the development of high-power communication systems, television equipment, radars, radio navigation systems, etc.

In their turn, electron vacuum devices are classified into two principal classes: *high-vacuum* (or simply vacuum) electron devices and *low-vacuum* (or gas-filled) electron devices. The former, in which electrons move in a highly evacuated space (the pressure is 10^{-7} mm Hg or less) feature high velocity of electron flow and are suitable for generation and amplification of signals in wide radio-frequency bands. In everyday life these devices are called radio valves. Gas-filled devices contain gas (the pressure is 10^{-3} mm Hg or more) and, therefore, the electrons moving in them collide very frequently

with gas molecules, thus causing ionization. As a result, not only electrons but ions as well take part in the operating process. Such devices are used in electric power equipment, particularly in AC converters (rectifiers), as well as in automatic and radio units where they, for the most part, perform the functions of automatic switching. Here, we shall discuss the operating principles of electron devices; the gas-filled devices are covered elsewhere.

Semiconductor devices have their own way of charge transfer and distribution inside the crystals of solid bodies under the action of electric and magnetic fields. In many instances such devices are not only capable of replacing radio valves but also to the benefit of the national economy have opened new branches of applied electronics. Of particular importance was the application of semiconductor devices in those installations which include dozens of thousands of active elements (for instance, electronic computers) since semiconductor devices made it possible to reduce their sizes (miniaturize) and increase the operational reliability of electronic equipment.

Molecular electronics is an advanced technology offering new ways in miniaturization of electronic equipment. It provides means for realizing radio circuits in a solid body, i.e., for the further development of semiconductor technology. Certain impurities (foreign atoms) introduced into a crystal make some crystal areas differ in their electric properties so that they may function as resistors, semiconductor devices, etc. It is precisely a combination of such areas that forms an electronic circuit in a semiconductor crystal. Devices developed on the basis of molecular electronics are extremely small in size and present a very important achievement in the field of microminiaturization of radio-electronic equipment.

Quantum electronics is based on the latest achievements of physics in the research of quantum phenomena occurring inside the atoms and molecules of a substance which may be in a solid, gaseous or liquid state. On the basis of these phenomena devices of different types such as lasers and masers are developed which can radiate electromagnetic waves in a variety of frequency bands. For instance, the generated waves may be about one micron long, i.e., in the region of visible (light) and infrared spectra. Such devices produce practically parallel light beams of high intensity making it possible to concentrate

a colossal amount of energy in small volumes. These devices find extensive application even in space technology.

Within the framework of this book we deal only with those branches of electronics which are directly associated with generation, amplification and conversion of radio-frequency and audio-frequency signals. The specific features of electronic computers and the equipment of the optical and adjacent bands are discussed in special technical literature. Therefore, the present chapter is devoted to electron vacuum and semiconductor devices and only briefly acquaints the reader with other problems of electronics. Arranging the material from the simple to the complex, we begin with electron devices.

8-2. MOTION OF ELECTRONS IN VACUUM.

CATHODES OF ELECTRON VALVES

The electron discovered at the very end of the last century presents so far an object of scientific research which reveals its new and ever new properties. It was found that the electron possesses a negative charge $e = 1.6 \times 10^{-19}$ coulomb (C) and has a rest mass * $m = 9 \times 10^{-28}$ gram (g).

Imagine a space between two plates (electrodes) in an evacuated glass envelope (Fig. 8-1). As will be recalled, to evacuate or create a vacuum in an envelope it is necessary to extract the air by special pumps. For the chemical absorption of gases which may be evolved inside the valve envelope as it operates, special absorbers (getters) are inserted into the valve, for example, barium is

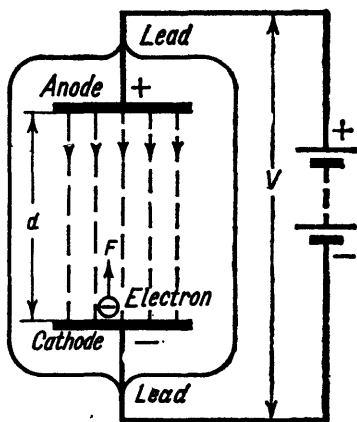


Fig. 8-1. Electrons in accelerating field

* As the velocity of the electron increases, its mass also increases; the velocities of electrons in electron vacuum devices, however, are significantly lower than the velocity of light and, therefore, the electron mass may be considered to be constant.

sprayed inside the envelope during the production process.

Imagine further that before evacuation the envelope is provided with conductors serving as plate leads which are sealed into the glass. Outside the envelope one conductor is connected to a positive terminal and the other to a negative terminal of a battery. In this case, the plate connected to the positive terminal of the battery is called the *anode*, and the plate connected to the negative terminal, the *cathode*.

Assume that an electron having a negligible initial velocity is released from the cathode surface. As a result, the electron falls under the influence of the electric field represented in Fig. 8-1 by lines of force directed from the anode to the cathode. A constant force attracts the electron to the anode.

This force can be easily computed. If we consider the field between the anode and cathode to be uniform, then its intensity

$$E = \frac{V}{d}$$

where d is the distance between the electrodes; V is the voltage across the electrodes.

The electric field acts on the electron with a force F equal (by Coulomb's law) to the product of its charge and the field intensity, i.e.,

$$F = eE.$$

we note here that this force is in a direction opposite to that of the electric field. As a consequence, the electron will start to move from the bottom upwards acquiring acceleration a which, according to Newton's second law, is

$$a = \frac{F}{m} = \frac{eE}{m} = \frac{eV}{md}$$

The motion will be a uniformly accelerated one. Consequently, when the electron is in motion, its velocity will increase as will its kinetic energy which attains the following value at the end of travel:

$$W = \frac{mv^2}{2}$$

where v is the velocity at the end of distance d .

The energy acquired by the electron as it moves in an accelerating field is equal to the work done by the electric field, i.e., the power expended by the battery. This work is the product of the force and the electron path length

$$W = Fd = eEd = eV \quad (8-1)$$

Equating the energy expended by the electric field to that gained by the electron, we may write the fundamental equation of electron motion in the electric field

$$eV = \frac{mv^2}{2}$$

hence, we find the velocity of the electron at the end of its path

$$v = \sqrt{2 \frac{e}{m} V}$$

Knowing that the ratio of the electron charge to the electron mass under the above specified conditions is a constant and expressing the velocity in practical units (km/s), we obtain

$$v \approx 600 \sqrt{V} \quad (8-2)$$

Thus, if voltage across the electrodes $V=100$ V, the electron velocity at the end of its path will be 6000 km/s, which, as we see, is much less than the velocity of light (300,000 km/s).

It is also easy to compute the time required for an electron to travel from the cathode to anode for a given distance d between them. It is well known that an average velocity of uniformly accelerated motion $v_{av}=v/2$, if the initial velocity is zero. Consequently, the time of electron travel called the electron transit time is

$$t = \frac{d}{v_{av}}$$

For instance, if in the above case $d=1$ cm (i.e., 10^{-5} km) the electron transit time

$$t = \frac{2d}{v} = \frac{2 \times 10^{-5}}{6 \times 10^3} = \frac{1}{3} \times 10^{-8} \text{ s}$$

i.e., about 0.003 μs or 3 ns (where ns is a nanosecond equal to 10^{-9} second).

Thus, the field which accelerates the electron does the work of increasing the kinetic energy of the electron. When the electron strikes the anode surface, its kinetic energy is converted into thermal energy since electron impacts heat the anode. Electron energy is measured in joules. Formula (8-1), however, makes it possible to express the energy of electrons in terms of another unit. As was shown, the electron energy is equal to the product of the electron charge e and voltage V under which action the electron was set into motion. Taking the electron charge equal to unity, we obtain the energy numerically equal to the number of volts. This unit of energy is called the *electron-volt* (eV). In the example given above, the electron on reaching the anode acquires the energy of 100 eV which amounts to $1.6 \times 10^{-19} \times 100 = 1.6 \times 10^{-17}$ joule (the electron charge $e = 1.6 \times 10^{-19}$ coulomb).

Let us consider the second case of the electron's motion in the same envelope. Assume that an electron being near the anode surface experienced an impact directed towards the cathode. Thus, the initial velocity of the electron is directed opposite to the force exerted on the electron by the electric field. In this case, the electric field proves to be not accelerating but retarding, so that the electron will be retarded uniformly. The energy gained by the electron at the instant of the initial impact will be expended as it moves in the retarding field, i.e., the electron will give up its energy to the electric field. A case is possible when the electron has lost all its energy before reaching the cathode. It will stop for a very brief moment to be accelerated back to the anode. As this takes place, the electron will be in a motion with constant acceleration.

The third typical case takes place when an electron is imparted an impact in a direction transverse to the electric field. The motion will not be linear since the electron will be deflected from its transverse path towards the anode. The degree of this deflection will depend both upon the initial velocity of the electron and the field intensity ($E = V/d$).

These are the simplest cases of interaction between an electron and the electric field. Interaction manifests itself at all times in the change of electron velocity and in the exchange of energy between the electron and the field produced by the anode battery. In an accelerating field the electron acquires

energy from the battery and in a retarding field it gives up energy to the battery. If the electron were moving not in vacuum but in a space filled with gas, then the above described processes would be disturbed due to collisions between the electron and gas molecules. Interaction between the electron and the magnetic field will be discussed somewhat later.

Let us now assume that close to the cathode electrons continuously appear and move to the anode under the action of an electric field. In this case, current starts to flow in the vacuum and in the battery circuit. Its magnitude is determined by the number of electrons reaching the anode each second. Thus, if about 7×10^{12} electrons transit between the electrodes each second, the current is 1 μ A. It is customary so far to indicate the direction of current flow as reverse to that of electron flow.

Conditions required to produce and maintain a current in a vacuum device can be provided by selecting a cathode which displays the property of electron emission from its surface. Such a cathode may be called an emitter (i.e., the electrode which emits electrons) and an anode may be referred to as a collector (i.e., the electrode which collects electrons). However, the terms "emitter" and "collector" are used primarily for semiconductor devices.

However, the cathode in the form of an ordinary metal plate shown in Fig. 8-1 cannot ensure noticeable continuous electron emission (at least under the action of ordinary anode voltages measured by dozens, hundreds and even thousands of volts). The point is that atoms in metals like in any solid body are tightly bound to one another to form a crystal lattice. Between the atoms of this lattice are gaps which appreciably exceed the electron size. Metals differ from dielectrics by the fact that their outer-shell electrons are loosely bound with the nucleus and are free to leave atoms. As this takes place, the lattice atoms become positive ions and the free electrons move about at random inside the metal. The total metal charge, however, remains equal to zero since the sum charges of ions and electrons are equal to each other. There are, of course, deviations from this average equilibrium state (called charge *fluctuations*) but they can be detected only with the aid of very sensitive instruments.

While moving freely from one atom to another, electrons cannot break loose from the metal surface because of the pre-

sence of the so-called "potential barrier", i.e., repulsion which tends to drive electrons back into the bulk of the metal.

The "barrier" is produced, firstly, by the electrons themselves which do break through the metal surface but are attracted back by the positive ions of the crystal lattice. A great number of escaping and returning electrons form, so to say, a layer of negative charges in the immediate (atomic distance) vicinity of the metal surface.

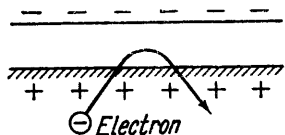


Fig. 8-2. Double electric layer on metal interface

The combination of ions on the metal surface and the electrons appearing above the surface is known as the *double electric layer* or the *dipole layer*. This is a narrow space one atom thick. The lower boundary of this layer is positive, and the upper one negative (Fig. 8-2). The action of the

double layer on each electron escaping through the metal surface is a retarding one: the force (the potential difference) acting on the electron is directed towards the metal surface and performs some negative work, i.e., decreases the kinetic energy of the electron.

Even if the electron energy were sufficient to overcome the action of the double layer, then there would be the force of attraction produced by the metal which loses negative electrons and, therefore, becomes positively charged. This force quickly decreases with the distance from the metal surface. Thus, the potential barrier which tends to prevent electron emission from the metal is formed actually by the double layer and by the force of metal attraction.

If the electron has broken through the barrier, it comes under the action of the anode attracting field and starts to accelerate uniformly towards the anode. A ball moving upward without friction under the action of an impact as shown in Fig. 8-3 may serve as the model of an electron as it moves through the potential barrier. The steep portion of the hump represents the action of the double electric layer and the upper portion with the decreasing steepness, the force of metal attraction.

It is quite clear that the hump can be overcome only by the ball which has acquired sufficient kinetic energy during impact. Accordingly, only those electrons which possess suf-

ficient energy may escape from the metal and perform work to break through the potential barrier. The minimum energy required for this to be done is called the *work function* which characterizes the metals emission properties. In fact, the less the work function, the greater the emission ability of a material.

At room temperature only a negligible number of electrons in any metal possess sufficient kinetic energy to perform the work function. Emission may be increased if some additional energy is given to the electrons in the metal.

The three most extensively used methods of increasing the energy of electrons in a conductor are: heating of the conductor, which produces *thermionic emission*; exposure of the conductor's surface to light causing *photoelectric emission*; and bombardment of the conductor's surface with external electrons which have high velocities causing *secondary electron emission*. Of these three types *thermionic emission* is the one mostly used in radio valves. In these valves the cathode is heated by an electric current as is the filament in incandescent lamps. The cathode is generally heated from an independent supply source.

Raising the metal's temperature increases the average electron flow energy in the bulk of the metal and a considerable number of electrons attain sufficient velocity to escape from the metal into the surrounding free space. Fine metals produce noticeable emission only at temperatures measured by several thousands of degrees (Centigrade). Such heat can be withstood only by refractory metals such as tungsten which is extensively used for the cathodes in radio valves.

Let us consider a *diode* (a two-electrode valve) used for rectifying alternating current, signal detection and other functions in radio systems. This is the simplest example of a radio valve (a thermionic vacuum device).

The diode consists of an anode and cathode (two electrodes) enclosed in a glass evacuated envelope (Fig. 8-4a). The anode in this example is made in the form of a metallic cylinder arranged horizontally and having a lead which protrudes thro-

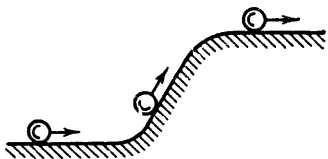


Fig. 8-3. Model illustrating electron motion through potential barrier

ugh the top of the glass envelope. A cylinder type of anode is more convenient than a plane one since the former encloses completely the cathode, although a great variety of anode shapes exist. The cathode in the form of a filament is positioned on the cylinder axis and fitted with two leads for connection to a battery. It is customary to designate valve electrodes as shown in the schematic diagram in Fig. 8-4b.

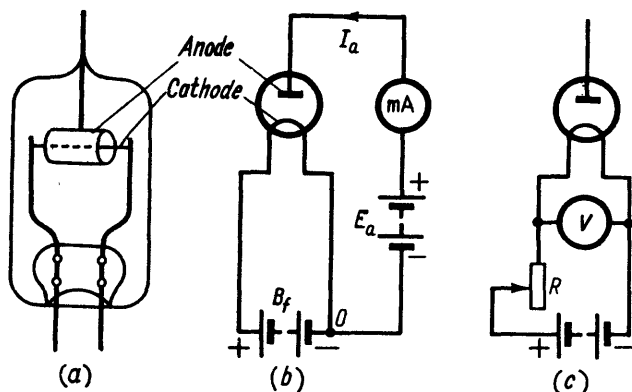


Fig. 8-4. Diode design and schematic diagrams

(a) simplest diode; (b) diode connected into circuit; (c) filament-voltage adjustment

Both leads of the cathode (filament) are connected to an independent filament battery B_f . The current supplied by this battery heats the filament to the temperature at which sufficient electron emission is obtained. If it is necessary to set a predetermined filament temperature, the filament leads are connected to a filament voltmeter V and a rheostat R (see the schematic diagram in Fig. 8-4c).

The anode battery E_a is connected between the anode and one of the cathode leads with the right polarity to ensure attraction of electrons to the anode. Anode current is measured with a milliammeter mA connected into the anode circuit. The current is equal to the sum of electron charges reaching the anode in one second. The anode current circuit runs from the positive terminal of the anode battery E_a to the anode, then from the anode to cathode through an evacuated space and further through the cathode to the common point O of the two

batteries and finally to the negative terminal of the battery E_a (Fig. 8-4b). The direction of anode current I_a is shown opposite to that of electron flow.

If the polarity of the anode battery is changed so that the anode becomes negative with respect to the cathode, the electric field would be not accelerating but retarding one and the electron flow to the anode would cease. This fact makes it possible for diodes to be used as alternating current rectifiers.

Let us now direct our attention to *cathode design*. Being heated to a specified temperature, each metal, including tungsten, emits a certain (average) number of electrons from each square millimeter of its area. Consequently, the cathode must have the sizes which could ensure the required emission current from its surface. The greater the anode current we want to obtain, the greater must be the cathode size and, consequently, the higher power required for its heating (necessary but not useful power consumption).

In this respect pure tungsten is not economical since it has to be heated to such a high temperature (not less than 2200°C) that the metal starts to evaporate and, therefore, the cathode serves only for short periods (average up to 1000 h). Only in very high power valves provision is made for cathode replacement allowing for tungsten cathodes to be extensively used in such valves.

To reduce cathode operating temperatures, still retaining adequate electron emission, it is necessary to assist the electrons to escape from the metal. For this purpose, the cathode is coated with a layer of another metal of one atom thick. The contact-potential difference between the two metals partially compensates the retarding action of the base-metal surface, thus facilitating electron escape. Therefore, the cathode coating process is generally called cathode *activation*.

Among such "activators" is thorium which is used for coating the specially carbonized surface of a tungsten filament. The operating temperatures of these carbonized cathodes do not exceed 1700°C (a yellow heat), and sometimes such cathodes are used in valves designed for medium-power transmitters. A valve with the carbonized cathode may fail due to the destruction of its activation coating rather than in the result of filament burnout. In such cases it is said that a valve "lost its emission".

Oxides of barium and strontium are other types of activators, but in this case nickel is used as the base metal. The activation layer is formed during cathode heat treatment. Oxide-coated cathodes operate at temperatures of about 800°C (a cherry-red heat). They have some excess of the activation metal in the oxide and, therefore, are less sensitive to overheating and their service life is much longer (for instance, 5000 h) than that of thoriated cathodes. In addition, they are capable of giving increased emission during short periods of time at high anode voltages (pulsed operation).

Valves with such *directly heated* cathodes, however, are hardly encountered in broadcast receivers supplied from AC mains. The fact is that if a directly heated cathode (a filament) is supplied with alternating current, the common point *O* (see Fig. 8-4*b*) proves to be alternately under the positive and negative potentials with respect to the filament centre. As a consequence, the anode voltage varies (pulsates), thus causing a low-frequency noise or hum in the receiver's loudspeaker.

In receiving and television equipment the valves with cathodes of other type are more widely used. Although being supplied with alternating current, these cathodes yet display a constant potential with respect to the anode (equipotential cathodes). In this cathode design the filament serves only to heat the cathode itself separated from it by an insulation, rather than to ensure electron emission. Such cathodes are called *heater* cathodes or *indirectly heated* cathodes.

The design of an indirectly heated cathode is illustrated in Fig. 8-5. The cathode itself is the metal cylinder *1* whose surface is coated with a layer *2* of oxides containing barium. The cylinder encloses a filament (heater) *3* in the form of twisted loop, which is coated with a heat-resistant layer of alundum *4* (an insulating material based on aluminium oxide). In another type of indirectly heated cathodes the heater is made in the form of a flat spiral and the cathode is a low cylinder with the top end closed by an oxide-coated cover.

A schematic diagram of a diode fitted with the indirectly heated cathode and the heater supplied from an AC mains is shown in Fig. 8-6. In some instances, distinct from this figure, the cathode is connected to one of the filament ends. In any case, however, the filament (heater) circuit and anode circuit are separated from one another so that the filament

alternating voltage does not affect the anode current. Furthermore, indirectly heated cathodes have a considerable thermal inertia owing to their relatively large mass and, therefore, maintain a constant average temperature in spite of the fact that the heater is supplied with alternating current.

Heating the cathode by the filament through an insulation, of course, requires an additional power. The problem of power economy is not so pertinent when it is supplied from AC mains as compared with battery supply. It takes several seconds for massive indirectly heated cathodes to reach the operating temperature when heater current is turned on. In circuits where the valve

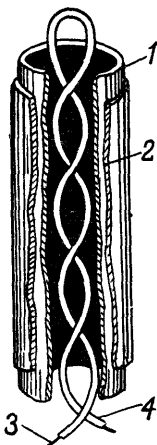


Fig. 8-5. Indirectly-heated cathode

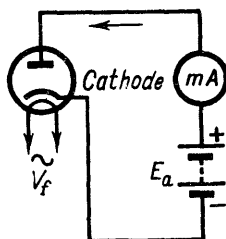


Fig. 8-6. Connection of diode with indirectly-heated cathode

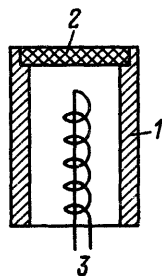


Fig. 8-7. "Impregnated" cathode

cathode is not connected to the heater, a voltage sometimes is applied across them. This voltage should not exceed a hundred volts as at higher voltages the alundum insulation breaks down.

High-grade valves with indirectly heated cathodes have a service life of 5000 h and more.

Power amplifier valves designed for transmitters employ *impregnated* cathodes. In these cathodes (Fig. 8-7) a disc cover 2 made of porous tungsten is attached to a molybdenum cylinder 1. The tungsten cover is impregnated with a suitable molten barium compound, barium aluminate. Such a cathode is heated to a temperature of about 1000°C and its service life under severe operating conditions exceeds 5000 h. The heater 3 is enclosed inside the cylinder.

The cathode types described above are only a few of a great variety of cathode designs used in modern electron valves.

8-3. DIODES

Let us consider in more detail the processes occurring in an anode circuit of a diode (as shown in Fig. 8-4*b* or Fig. 8-6). Assume first that no anode current flows in the circuit, i.e., the anode is shorted to the cathode. Electrons emitted by the cathode remain in the space near the cathode since their further motion in vacuum is prevented by the repelling effect of electrons which have escaped earlier. Thus, a cloud of electrons is formed around the cathode. The combination of charges of these electrons is generally referred to as a *space charge*.

The density of the space charge, i.e., the mean number of electrons per unit volume, increases with the increasing temperature of the cathode. At a given (say, normal) temperature the density of the space charge remains constant since the cloud of electrons retards the motion of newly emitted electrons returning them back to the cathode. Separate electrons which have escaped from the cathode at high initial velocities may reach the anode and induce some initial current in its circuit (i.e., the current flowing in the absence of anode voltage). This current, however, is not sufficient for practical use.

An electric field arises whenever a battery (or any other source of DC voltage), producing a positive potential with respect to the cathode, is connected into the anode circuit. Some lines of force of this electric field may be represented as terminating in the electrons of the space charge, and these electrons are set in accelerated motion towards the anode (Fig. 8-8). As a consequence, anode current arises. This current increases and the density of the space charge decreases with the increasing anode voltage. When the anode voltage is high enough, the retarding field disappears and all the electrons emitted from the cathode surface enter immediately the accelerating field which causes them to move towards the anode and any electron reaches it. A further increase in the anode voltage produces no further increase in the anode current. The condition in which all of the cathode emission is used up is called *saturation*. Under these conditions the anode current

I_a equals the total emission current I_e and may be called the *saturation current*.

If cathode temperature is increased (of course, within permissible limits), the density of the space charge will also rise due to an increase in energy of electrons moving inside the

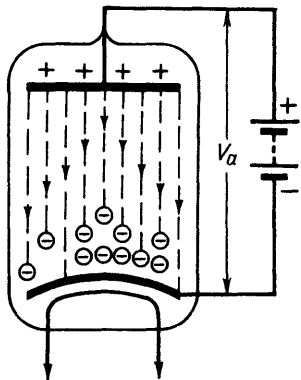


Fig. 8-8. Interaction between space charge and anode field

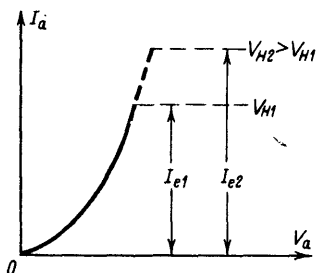


Fig. 8-9. Theoretical relationship between anode current and anode voltage (for two filament voltages)

metal. In this case, saturation will be reached at a higher anode voltage, and the saturation current I_e will rise. The theoretical relationship between the anode current I_a and anode voltage V_a is shown graphically in Fig. 8-9 for two filament voltages V_{H1} and V_{H2} . The anode voltage at which saturation current is reached may be called the *saturation voltage*. It is different for different filament (heater) temperatures.

It is common practice, however, to operate diodes under conditions when anode current is limited by a space charge.

As mentioned above, the latter tends to prevent the emitted electrons from moving towards the anode driving them back to the cathode. We may say that the space charge presents a

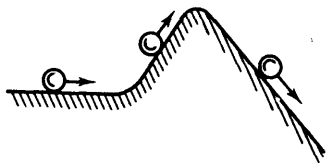


Fig. 8-10. Model of space-charge retarding action

potential barrier on the electron path. A hump with a negative gradient along which a ball rolls down after overcoming the retarding field produced by the space charge (Fig. 8-10) may serve as a mechanical model of the flow of the electrons which have escaped from the cathode. When approaching the anode, the electron gains its kinetic energy similarly to the ball moving down the hump.

The relationship between the anode current and the anode voltage, which is represented by the theoretical *anode characteristic* of a diode (diode characteristic) in Fig. 8-9 is *curvilinear* in the ascending portion. This means that anode current is not proportional to anode voltage. To explain this fact, let us imagine that the diode in Fig. 8-8 is cross-cut by a plane parallel to the anode. Then an electron current

$$i = \rho v$$

flows through each square centimeter of the plane, where ρ is the total charge of electrons in 1 cm^3 at the cross-section level, whereas v is the velocity of electron flow (cm/s) at the same level. The density of charge is proportional to anode voltage as it is determined by the number of electrons captured by the anode field lines of force. The velocity of electron flow is, in its turn, proportional to the square root of anode voltage as shown in formula (8-2). The closer the cross section to the anode, the greater the velocity of electrons. Consequently, the anode current taken from the entire cross-section area s at the total voltage V_a is

$$i_a = spv = kV_a \sqrt{V_a} = kV_a^{3/2} \quad (8-3)$$

This *three-halves power law* describes the ascending portion of the theoretical diode characteristic (see Fig. 8-9). The proportionality factor k is determined by diode construction.

The upper portion of the theoretical characteristic of a diode is horizontal. The characteristic has an abrupt transition from the space-charge-limited conditions to saturation operation in which saturation current remains essentially constant.

Diode characteristics may be taken experimentally by using a measurement set whose schematic diagram is shown in Fig. 8-11. The filament voltage V_H is maintained constant in measuring each characteristic curve. The anode voltage

V_a is varied with a potentiometer and readings are taken against the voltmeter; the anode current I_a is measured with the aid of the milliammeter.

The measurement characteristics shown in Fig. 8-12 differ significantly from the theoretical one while remaining similar to it in general.

First, the actual characteristics start in the region of negative anode voltages. This is explained by high-energy elect-

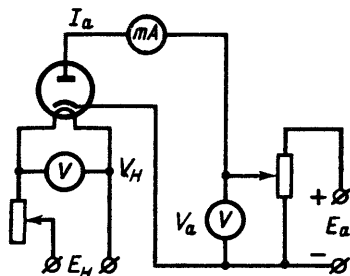


Fig. 8-11. Set-up for measuring the diode characteristics

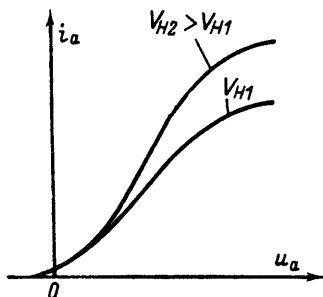


Fig. 8-12. Measured characteristic curves of diode

ons emitted from the cathode as well as by the contact-potential difference between the anode and cathode materials. This initial anode current is taken into consideration only in analyzing the diode operation under small-signal conditions (fractions of a volt).

Second, the lower segment of the actual characteristic curve in the region of positive anode voltages (the *lower knee*) rises steeper than in the theoretical curve since the theoretical relationship does not account for the initial velocity of electrons and the contact-potential difference.

The major portion of the characteristic curve is practically linear. Its departure from the three-halves power law is explained by a number of physical and constructional reasons such as nonuniform heating of the cathode, different electron velocities in the same cross section, anode-field effects on the emission process, etc. For instance, an increase in the energy of emitted electrons with the increasing cathode temperature causes characteristic-curve discrepancy even on their linear segment.

Finally, the transition to the saturation segment in the actual characteristic curve is found to be gradual (the *upper knee*) rather than abrupt as in the theoretical curve. Saturation cannot be too abrupt since it occurs earlier at the less heated portions of the cathode (for instance, near its attachment points) than for the more heated portions. After the upper knee, saturation current remains constant with the increasing anode voltage only for valves having a pure-tungsten cathode, whereas for diodes with oxide-coated cathodes it continues to rise. One of the reasons for this rise is the Schottky effect which consists in an increase of emission current when the anode electric field acts directly on the cathode, i.e., when there is no space charge between the cathode and anode. Another reason of the anode current rise, typical of oxide-coated cathodes, is additional cathode heating by the anode current itself. Because of the latter effect, the point of transition from space-charge-limited to saturation conditions cannot be determined from the characteristic curve.

A circuit incorporating a diode does not obey Ohm's law for the following reasons. First, the circuit current stops to flow when the sign of anode voltage is reversed. Second, the current varies non-proportionally with anode voltage even in the region of positive voltages. The devices characterized by a non-linear relationship between voltage and current are called *non-linear devices*. Thus, the diode is also a non-linear device.

Resistance (or its reciprocal, i.e., conductance) is the only quantity characterizing the relation between current in and voltage across any conductor connected into the DC circuit. Resistance may be used as a parameter of a linear circuit. A concept of parameters may be also introduced for non-linear devices and, in particular, for diodes. The *diode internal resistance* may serve as such a parameter. Voltage between the anode and the cathode and the current flowing through the space between these two electrodes may be measured with a voltmeter and ammeter, respectively. The ratio of voltage to current gives the direct-current internal resistance of the diode. However, from the diode characteristics discussed above one may make sure that this resistance is not a constant that is typical of a non-linear device. In fact, at negative anode voltages anode current is zero. Hence, in this region the diode

DC resistance equals infinity (and its conductance is zero). At positive anode voltages DC resistance will be different for different points of the characteristic curve due to its nonlinearity.

The DC resistance for the point *A* on the characteristic curve shown in Fig. 8-13*a* is

$$R_0 = \frac{V_{a1}}{I_{a1}}$$

i.e., is proportional to the cotangent of the angle formed by the straight line *AO* with the horizontal axis. But for other points of the characteristic curve a similar angle will be dif-

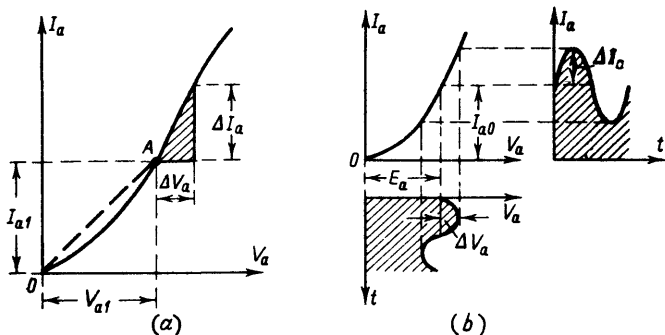


Fig. 8-13. Concept of DC and AC resistances of diode

(a) plotting of characteristic triangle; (b) relationship between AC current and AC voltage

ferent and, consequently, the DC resistance will also be different.

Possible also are such cases when an alternating voltage ΔV_a is applied to the anode of the diode in addition to the DC voltage E_a . In Fig. 8-13*b* this alternating voltage is shown along the time t which is arranged vertically downwards. In this case, pulsating current which contains both direct (I_{a0}) and alternating (ΔI_a) current components flows through the anode circuit. The alternating current is shown in Fig. 8-13*b* as a function of time. To find the relation between anode voltage ΔV_a and anode current ΔI_a , one should know the resistance R_i offered by the diode to alternating current. This AC resistance is determined within the operating portion

of the diode characteristic, i.e., the portion corresponding to the applied alternating voltage.

$$R_i = \frac{\Delta V_a}{\Delta I_a} \quad (8-4)$$

The *diode AC resistance* (or in other words the *dynamic resistance*) is the ratio of the anode-voltage increment to the appropriate increment of anode current. In formula (8-4) the symbol Δ stands for increment.

The anode-current time diagram shown in Fig. 8-13*b* is obtained by projecting the anode-voltage time diagram on the diode characteristic curve. Because of the nonlinear diode characteristic, the curve representing the current as a function of time may prove to be distorted when compared with the voltage-via-time curve. This type of *distortions* is called *non-linear*.

Since diode AC resistance is different for different portions of the characteristic curve, one may only speak about an average value of AC resistance for this or that portion of finite length. The major portion of an actual characteristic curve, however, is practically linear so that diode dynamic resistance remains approximately constant within this portion of the curve. Therefore, this value is also used as the diode parameter.

It is not necessary, of course, to plot a time diagram similar to that shown in Fig. 8-13*b* in order to determine the diode dynamic resistance. For this, one may use the plot shown in Fig. 8-13*a* and select on the diode characteristic curve the portion between the voltages V_{a1} and $V_{a1} + \Delta V_a$, to which the currents I_{a1} and $I_{a1} + \Delta I_a$ correspond. The selected portion and the segments representing the voltage and current increments form a triangle whose sides are equal to ΔV_a and ΔI_a , respectively. On measuring their length and taking into account the scale factors of the coordinate axes, we may find the approximate value of the diode dynamic resistance at the centre of the selected portion. Thus, we may write that

$$R_i = \frac{\Delta V_a}{\Delta I_a}$$

For example, with $\Delta V_a = 8$ V the current $\Delta I_a = 32$ mA = 32×10^{-3} A, hence we obtain

$$R_i = \frac{8}{32 \times 10^{-3}} = 250 \text{ ohms}$$

The AC resistance for the linear portion of the diode characteristic curve is less than the DC resistance which can be easily shown from Fig. 8-13.

The relationship between voltage and current increments can be also estimated with the aid of another parameter which is known as the *slope of the diode characteristic curve* and corresponds to the diode AC conductance. The slope is the reciprocal of the diode AC resistance, i.e.,

$$S = \frac{1}{R_i} = \frac{\Delta I_a}{\Delta V_a} \quad (8-5)$$

If AC resistance is measured in ohms, i.e., in volt per ampere, then the slope or diode conductance would be measured in ampere per volt, i.e., in siemens (S). In practice, however, it is convenient to measure conductance of low-power diodes in milliamperes per volt (mA/V), i.e., in millisiemens (mS).

The diode conductance is defined as the ratio of a slight increase in anode current to the slight increase in anode voltage that caused it. Numerically, diode conductance indicates by how many milliamperes the anode current will rise if anode voltage is increased by 1 V. Thus, if $R_i = 250$ ohms, the diode conductance or slope

$$S = \frac{1}{250} \text{ A/V} = 4 \text{ mA/V (4 mS)}$$

Consequently, diode conductance cannot be an independent diode parameter since it is determined by the diode AC resistance. The diode conductance can be also determined from the characteristic triangle shown in Fig. 8-13a. Just like the AC resistance of a diode, its conductance is different for different portions of the diode characteristic curve. In the region of negative anode voltages it is zero. The conductance of a diode corresponding to the linear portion of its characteristic curve within which diode conductance remains almost constant and equals its maximum value is generally used as a diode parameter.

In practical calculations, however, an *idealized diode characteristic* in the form of a straight line starting from the origin of coordinates to the region of positive anode voltages can sometimes provide satisfactory results. It is clear that such a characteristic gives only one value of the parameter R_i (or S). In the region of negative voltages, however, AC resistance is equal to infinity (diode conductance is zero). Consequently, the diode having such a characteristic curve still remains a non-linear device which conducts electric current only in one direction.

When developing ultrashort-wave equipment, one should take into account, apart from AC resistance of a diode, its "parasitic" parameter — the *anode-cathode capacitance* C_{ak} . Such capacitance is inevitably present in any diode as a result of the mutual geometric arrangement of its two electrodes together with their leads. At the same time this capacitance is undesirable since its susceptance ωC_{ak} is low at very high frequencies, so that it may appreciably shunt the diode, thus decreasing its rectifying effect. In the ultrashort-wave diodes (operating, for instance, at meter waves), the capacitance C_{ak} is equal to several picofarads. In kenotrons (the diodes designed for rectifiers) this parameter is negligible. Another diode parameter which is very important for desired operating conditions is the *anode maximum dissipation* $P_{a \max}$. As we know, each electron striking the anode gives up energy which is equal to the product of the charge e and voltage V_a and is converted into heat. If N electrons reach the anode every second, the total energy given up by these electrons is equal to the power expended for anode heating, i.e.,

$$P_a = NeV_a$$

where Ne is the quantity of electricity delivered to the anode every second, i.e., the magnitude of anode current, I_a .

Power P_a may be increased only to a certain limit, i.e., until the anode heating proves to be harmful to the valve (the anode material evaporates, evolves gases and even produces additional cathode heating). It is precisely this maximum power which is used as a specification parameter of diodes designed for rectifiers. To avoid damage to the anode in operation, it is important that $P_a < P_{a \max}$. In low-power diodes

employed in radio receivers there is no need to take into account anode heating as it is always below the permissible level.

To increase the maximum anode dissipation, special design measures are generally taken to keep anodes cool. For instance, they are made of blackened nickel providing good heat dissipation as well as cooling ribs are attached to anodes which increase their area. Forced air or liquid cooling is commonly used for high-power valves.

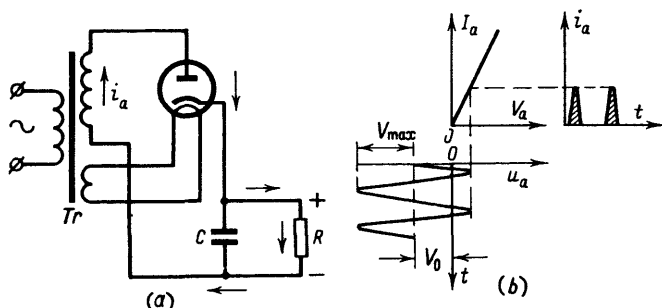


Fig. 8-14. Half-wave kenotron rectifier
(a) schematic diagram; (b) operating principle

The *peak inverse anode voltage*, $V_{p\ inv}$, is yet another operating parameter for high-voltage vacuum devices. As we shall see later from Fig. 8-14b illustrating the operation of a simple diode (kenotron) rectifier, the maximum anode voltage with respect to the cathode equals the sum of the rectified voltage V_0 and the alternating voltage V_{max} being rectified. This sum voltage is negative (with respect to the cathode). It must be less than the voltage capable of breaking down insulation between the leads, thus causing the diode failure. Consequently, $V_{p\ inv}$ should be chosen much less (by 2 or 3 times) than the breakdown voltage.

The maximum pulse current and the maximum rectified current are also used as limiting parameters of diodes (kenotrons). These parameters are determined by the emission properties of cathodes.

An important operating parameter of a valve is service life which depends, obviously, not only upon the long-term stability of cathode properties. If during valve testing (after operat-

ion or storage) one of the above parameters proves worse than its limit value, the valve is rejected. Therefore, service life of an electron device is the length of operation time during which the device provides specified performance under operating conditions until one of its major parameters falls beyond the specified limits or the device becomes unsuitable for normal operation due to other reasons.

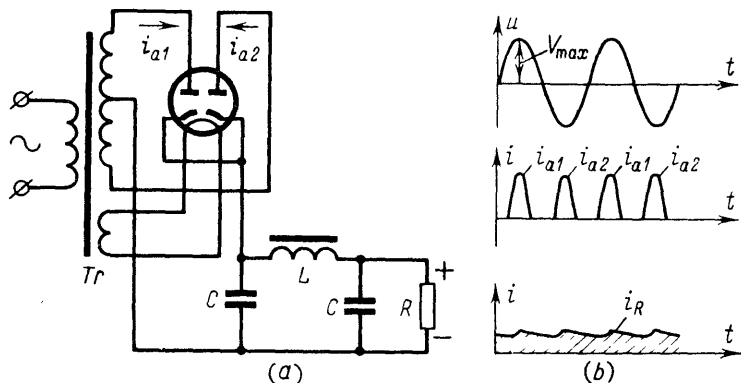


Fig. 8-15. Full-wave rectifier
(a) schematic diagram; (b) operating principle

As an example of a diode (kenotron) application, let us consider the schematic diagram of a simple rectifier used to supply the load resistor R (Fig. 8-14a) with direct (rectified) current. The rectifier is supplied through a transformer Tr whose primary winding is connected to AC mains. One of the secondary windings of the transformer supplies the kenotron filament, and the other produces alternating voltage in the anode circuit. The kenotron conducts only when anode voltage is positive. Under these conditions the current flows inside the kenotron from the anode to cathode and then through resistor R . Consequently, current pulses will pass through resistor R and the kenotron, thus producing a pulsating but not DC voltage across the load resistor.

A capacitor C connected in parallel to the resistor serves as a simple filter so that only direct current can flow through the load resistor. Thus, idealizing the filter performance,

we may say that a negative DC voltage developed across the load resistor is applied across the kenotron in addition to the AC voltage from the transformer. This simplified operation of the rectifier is illustrated by diagrams shown in Fig. 8-14*b* in which the kenotron characteristic in the region of positive anode voltages is represented by a straight line.

It should be noted that a current pulse through the kenotron lasts for less than a half-period since the current can pass only during time intervals when positive voltage from the transformer exceeds the direct voltage V_0 developed across the load resistor. During a negative half-cycle the voltage from the transformer sums up with that across the load as indicated above. Useful (rectified) voltage is equal to the product of the direct-current component and the load resistance. Such a rectifier which conducts only during one half-cycle of supply voltage is called the *half-wave rectifier*.

A *full-wave rectifier circuit* requiring two kenotrons or one twin kenotron (Fig. 8-15*a*) provides better performance. In this circuit both diodes operate alternately on the common resistor R . A lead to the load resistor is taken from the centre tap of the secondary winding of a step-up transformer. Diagrams of voltages across and currents in the anode circuits and resistor R are shown in Fig. 8-15*b*. Kenotron anodes are supplied from two different halves of the secondary winding. When voltage is positive for the first anode, it is negative for the second one and vice versa. Current pulses follow through the diode (its two halves) during some portion of each half-cycle, producing voltage pulses across load resistor R , which always have the same polarity.

The advantage of the full-wave rectification circuit is in that the current pulses follow one another twice the rate in the half-wave circuit. This aids filtering. To improve filtering, a lower-frequency filter in the form of Π -section (with choke L) is used in the circuit (see Fig. 8-15*a*).

The vacuum diode is the simplest electron device. It falls into the category of *passive* devices as it cannot amplify signal or maintain oscillations. Much more complicated and important functions are performed in radio systems by *active* devices among which a triode (a three-electrode valve) was historically the first.

8-4. TRIODES

Let us imagine a vacuum device with a flat emitting cathode, flat anode and the third electrode, namely, a wire grid located between the anode and cathode all contained in vacuum (Fig. 8-16a). Let us now denote the voltage across the anode with respect to the cathode by V_a , the voltage applied to the grid with respect to the cathode by V_g . The purpose of the grid is to change significantly the anode cur-

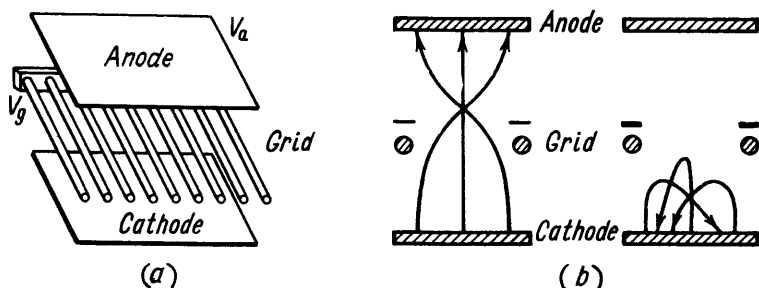


Fig. 8-16. Function of grid

(a) schematic representation of electrodes; (b) electron paths for different negative grid voltages

rent I_a by varying slightly the voltage V_g . In other words, the *grid* is an electrode *controlling the anode current*.

Such a three-electrode valve is called the *triode*. It may be used for signal amplification and generation.

Electrons moving from the cathode to the anode generally pass between the grid wires but the number of electrons passing through the grid each second depends essentially upon the grid voltage. If this voltage is negative (with respect to the cathode), the grid has a retarding effect on the electrons, thus producing a repelling barrier on their path. A negative voltage may be applied to the grid such that the accelerating action of anode field will disappear near the cathode. With a further increase in grid negative voltage the field in the vicinity of the cathode becomes a retarding one, so that anode current ceases. It is customary to say in this case that "*cut-off*" voltage is applied to the grid.

If positive voltage (with respect to the cathode) is applied to the grid, the accelerating field intensity in the space be-

ween the grid and cathode increases, thus causing an increase in anode current. Consequently, anode current can be varied from zero up to some maximum value by changing grid voltage.

Figure 8-16*b* shows the trajectories of electrons emitted from the zone of a space charge at right angles to the cathode surface when anode voltage is positive and grid voltage is negative. The left drawing corresponds to a slight negative voltage at the grid (only two adjacent wires of the grid are shown). Electrons move initially along linear paths but then they are deflected by the negatively charged grid wires. The closer the initial path of an electron to the grid wire the greater the deflection. In some region behind the grid the electron paths intersect or, as it is customary to say, the electron flow is focused and then the flow approaches the anode in the form of a divergent beam. The right drawing corresponds to an appreciable negative grid voltage. Under these conditions, electrons cannot clear the barrier produced by the grid electric field and return back to the cathode. As a consequence, no anode current flows in the circuit.

With a positive grid voltage not only is the anode current increased but some part of the electrons moving from the cathode may be intercepted by the grid. Thus, current will exist not only in the anode circuit but in the grid circuit also (*grid current*). In general, the grid current I_g serves no useful purpose.

In actual triodes the grid is given the form of a coil and an anode, the form of a high or low circular cylinder. The physical processes occurring in such triodes are similar to those described above. However, in the cylinder-type designs the electric-field lines of force and electron paths are along cylinder radii.

Twin triodes are in wide use. In such a design both triodes are enclosed in one envelope and can be used either in the same stage or perform different functions. For receivers it is also convenient to use triode and twin diode combinations enclosed in a common envelope.

As shown above, currents in anode and grid circuits of a triode depend upon the anode and grid voltages (provided the filament voltage is set normal for the valve used, i. e., it ensures a required cathode emission). A quantitative esti-

mate of relationships between these voltages may be presented graphically by *triode characteristic curves*.

The main characteristic of triode is the graphical relationship between anode current and grid voltage when anode voltage is maintained constant. This is a grid characteristic of anode current or in short an *anode-grid characteristic*. The anode- and grid-current circuits as well as the filament circuit of a typical triode are shown in Fig. 8-17. Milliamme-

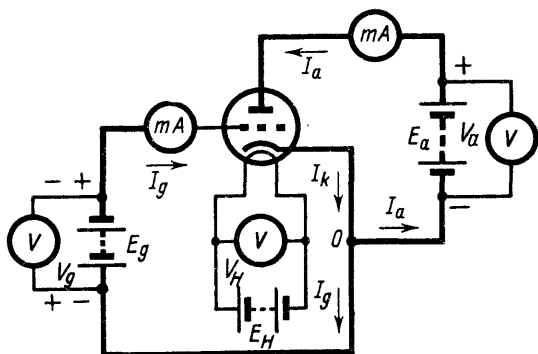


Fig. 8-17. Set-up for measuring the voltages and currents in anode and grid circuits

ters *mA* measure the anode current I_a and grid current I_g . The anode and grid voltages as well as filament voltage are measured by voltmeters *V*. It is evident that the cathode current I_k is the sum of anode and grid currents ($I_k = I_a + I_g$) which branch off at the *common point O*. Therefore, a separate instrument to measure the cathode current is not required. For clarity, anode and grid circuits are shown by thick lines.

On setting the normal filament voltage V_H and anode voltage V_a as specified in the valve certificate, let us vary the grid voltage V_g from negative through zero to positive values. In principle, this may be achieved by connecting the grid lead to the leads of different battery elements. It is more convenient, however, to use a voltage divider (a potentiometer) which is not shown in Fig. 8-17 for the sake of diagram simplicity (see Fig. 8-11). We shall record anode current for each grid-voltage setting.

No anode current I_a flows when an appreciable negative voltage is applied to the grid since electrons are repelled by the grid back to the cathode. As we decrease gradually grid negative voltage, a current appears in the anode circuit; at first it will rise slowly and then quicker. This shows that the anode positive voltage can overcome the action of grid negative voltage if the grid voltage is made much less than anode voltage in its absolute value.

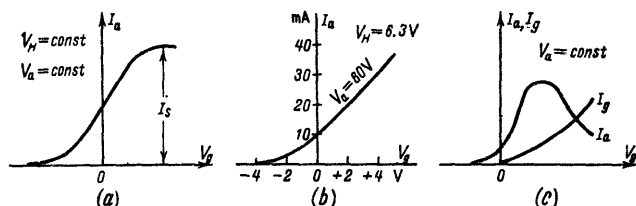


Fig. 8-18. Triode characteristic curves

(a) anode-grid characteristic for valve with tungsten cathode; (b) anode-grid characteristic for low-power amplifying valve with heater cathode; (c) anode and grid currents as functions of grid voltage

On reaching zero grid voltage and after reversing the polarity of the grid battery E_g , we further increase grid voltage, but now in the positive region. Under these conditions, we shall observe a further increase in anode current. If the valve under test had a pure tungsten cathode, the anode current increase would slow down with the increasing positive grid voltage and finally the anode current would cease. This would correspond to saturation conditions.

The plot (Fig. 8-18a) representing the resultant relationship is called the anode-grid characteristic. The saturation current I_s (at rated filament voltage) characterizes the properties of a valve having a tungsten cathode. The steep (practically linear) portion of the anode-grid characteristic is its most important (operating) portion.

For a valve having an activated (say, oxide-coated) cathode saturation current cannot be specified since emission current continues to rise at the expense of additional heating of the activated coating by anode current. Therefore, the characteristic curve is terminated on reaching the anode current which is still safe for the cathode. An approximate anode-

grid characteristic of a low-power triode having a heater-type cathode is shown in Fig. 8-18*b*. Similarly to the previous characteristic curve (Fig. 8-18*a*), it has a lower knee in the region of grid negative voltages.

The grid current plotted as a function of grid voltage for a fixed anode voltage is the second characteristic of a triode, which is called the grid characteristic. As we know, current appears in the grid circuit only when positive voltages are applied to the grid so that grid wires may partially intercept electrons. The current which is, of course, opposite in direction to the electron flow passes through interconnecting wires from the cathode to the grid and inside the valve from the grid to the cathode (see Fig. 8-17). With small grid positive voltages, grid current is insignificant. In fact, it amounts to a fraction of a per cent or several per cent of anode current. As the positive voltage applied to the grid increases, the number of the electrons intercepted by it also increases, and grid current becomes comparable with anode current. With a comparatively low anode voltage redistribution of electron flow is possible. This redistribution leads to an increase in grid current at the expense of a decrease in anode current when positive voltage applied to the grid exceeds the anode voltage. In Fig. 8-18*c* the grid and anode currents are shown over a wide range of grid-voltage values with the anode voltage being maintained at a low constant level. It should be emphasized, however, that wide-range variations of grid voltage such as required in taking the above characteristics may damage the valve since grid leads are strongly heated by electron bombardment when grid current is too high. It might be well to point out once again that in most cases grid current serves no useful purpose in the radio circuits and power consumption due to grid heating is harmful.

Let us now consider how *anode-voltage variations* affect the processes occurring in the valve and valve characteristics. After taking the characteristic as shown, for instance, in Fig. 8-18*b*, we increase anode voltage and once again take the anode-grid characteristic. When anode voltage is increased, electrons are attracted with a much greater force to the anode so that they can easier overcome the retarding action of the negatively charged grid. Therefore, the anode-grid characteristic curve will start at an increased distance on the

left of the coordinate origin, i. e., at a greater negative grid voltage, and all anode current values will be higher than the previous ones at the same grid voltages.

In Fig. 8-19a anode-grid characteristic curves are shown for $V_a=60, 120, 180$ and 240 V. To a first approximation, all these characteristics differ only in their position with respect to the coordinate axes. This means that if the anode

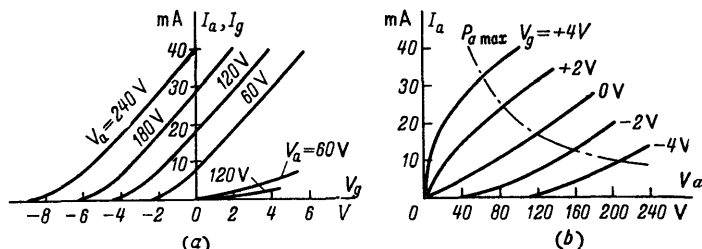


Fig. 8-19. Family of triode characteristic curves

(a) anode and grid currents as functions of grid voltage at different anode voltages;
(b) anode characteristic curves at different grid voltages

voltage is increased by equal steps then each succeeding characteristic can be obtained (at least approximately) by moving the previous one *to the left* by one and the same distance.

The curves plotted below in the same figure present grid characteristics taken for anode voltages $V_a=60$ and 120 V. No wonder that at higher anode voltages, grid current is less for the same grid voltages (since the force of anode attraction will be greater so that a smaller portion of electron flow will be intercepted by the grid).

A group of characteristic curves of the same type taken for different values of DC voltage is called a *family of characteristics*. Fig. 8-19a illustrated a family of anode-grid characteristics (as well as a family of grid characteristics) obtained for several anode voltages. Note that the major (steep) portions of anode-grid characteristics are essentially *straight lines*; the importance of this fact will be explained later.

Valve performance may be also estimated by using another type of characteristics representing relationships between the anode current and anode voltage for a given grid voltage.

Characteristics of this type are simply called *anode characteristics*.

Returning to the schematic diagram shown in Fig. 8-17 and setting the grid voltage equal to zero ($V_g=0$), we shall increase anode voltage from zero to high positive values. As a result, we shall obtain a characteristic curve which is in principle similar to the characteristic curve of a diode (Fig. 8-19*b*). In fact, when $V_a=0$, the current is zero since the anode does not attract electrons. This means that the characteristic curve will start from the coordinate origin and with a further increase in anode voltage will have a lower knee and then an almost straight-line portion.

When the grid is fed with a negative DC voltage ($V_g<0$), the anode characteristic curve is shifted somewhat to the right from the coordinate origin since the anode voltage must be increased to overcome the increased retarding action of the grid. The greater the negative grid voltage, the greater will be the shift of the characteristic (see Fig. 8-19*b*).

When a fixed positive potential is applied to the grid, the anode characteristic curve is on the left of the main one (which corresponds to $V_g=0$). Although such characteristic curves also start from the coordinate origin, their initial portions prove to be convex. This abrupt rise in anode current may be explained by the redistribution of electrons between the positive grid and the anode—with a very low anode voltage electrons which have passed through the grid enter the retarding field between the grid and the anode forming between them a *second space charge*. As the anode voltage is increased, the anode current rises sharply at the expense of the electrons taken from this space charge which disappears, and also because of the abrupt grid current decrease.

In Fig. 8-19*b* the valve operating region is limited by a dashed line which corresponds to a maximum anode temperature that can be tolerated, i. e., this line passes through the points corresponding to the *maximum anode dissipation* ($P_a=V_aI_a$). For the valve whose characteristic curves are presented in this figure the maximum anode dissipation $P_{a\text{ max}}$ amounts to 2 W. Instantaneous (pulse) power may exceed this value since the anode cools between pulses.

The characteristic curves discussed above are called *static characteristics* as they are taken with DC voltages applied to

the electrodes. These characteristic curves give a fairly complete idea about triode performance. It is desirable, however, to express this performance in figures. Such quantitative assessment is provided by the use of *triode parameters*.

A triode is designed primarily for signal amplification. Amplification in a triode is possible due to the fact that its grid is located between the anode and cathode, thus shielding partially the cathode from the direct action of the anode field. The field of the grid itself acts on the space charge without any hindrance. The valve parameter indicating how many times the grid voltage is more effective than the anode voltage in the controlling the anode current is called the *amplification factor* of a valve and is denoted usually by the Greek letter μ .

Increasing the grid voltage by ΔV_g volts we will observe the resultant increase in anode current. If we wish to restore the previous anode current value by decreasing anode voltage, we should decrease it by ΔV_a with ΔV_a being greater than ΔV_g . It is precisely the ratio of these voltages that serves as the valve amplification factor, i. e.,

$$\mu = \frac{\Delta V_a}{\Delta V_g} \quad (8-6)$$

when current I_a is held constant.

It is evident that $\mu > 1$ and is a dimensionless quantity. For instance, if the amplification factor $\mu = 25$, an increase of grid voltage by 0.4 V will cause the same increase in anode current which would be caused by an increase of anode voltage by $0.4 \times 25 = 10$ V.

Sometimes, it is more convenient to use the reciprocal of the amplification factor. This parameter is called the *grid-through* and is denoted by the letter D ,

$$D = \frac{\Delta V_g}{\Delta V_a} \quad (8-7)$$

(for $I_a = \text{constant}$).

Physically, the grid-through shows approximately which portion of anode field lines of force penetrates through the grid to the cathode. It is evident that the closer the grid wires are to each other, the less is the grid-through, i. e., the greater is the amplification factor. Different applications

require triodes with amplification factors ranging from a few units to a hundred.

The physical definition given above for the grid-through deals with the effect of the anode field upon the cathode. Cathode emission, however, produces not only anode current but grid current as well. The latter was not considered in defining the amplification factor μ . Therefore, the equality $D=1/\mu$ is accurate only for left portions of anode characteristic curves corresponding to zero grid current. But, if the grid current is small compared to the anode current, the error introduced when using the equality $D=1/\mu$ is insignificant.

Another triode parameter which characterizes the effect of the grid voltage on the anode current when the anode voltage is held constant is referred to as the *slope* of the anode-grid characteristic curve or, as it is customary to say, the *valve transconductance*. Transconductance indicates by how many times the anode current increases when the grid voltage is increased by 1 V. It is designated by the symbol S and is defined as

$$S = \frac{\Delta I_a}{\Delta V_g} \quad (8-8)$$

when the voltage V_a is held constant.

Consequently, from the physical point of view transconductance is measured in the same units as conductance. It presents, so to say, "mutual conductance" of the grid and anode circuits rather than conductance of any one valve circuit. It is common practice to measure transconductance in millisiemens (mS), i. e., in milliamperes per volt (not in amperes per volt). Assume, for instance, that an increase of 0.5 V in grid voltage has caused an increase of 5 mA in anode current, then transconductance will be equal to 10 mA/V.

Amplifier triodes used in receivers have transconductance from 1-2 to 30-40 mA/V. This parameter is also determined by the valve design—the larger the cathode emitting surface and the closer the grid to the cathode, the higher the transconductance. Needless to say that the transconductance of a triode (or diode) is specified for the linear (steep) portion of the characteristic curve. In most applications high transconductance of a valve is an advantage.

Anode current may be also varied by changing the anode voltage while maintaining the grid voltage constant. Having increased the anode voltage by ΔV_a we find that anode current has increased by ΔI_a . This relationship may be estimated quantitatively with the aid of another triode parameter which is called the AC *anode resistance* and is denoted by the symbol R_i . It is defined as

$$R_i = \frac{\Delta V_a}{\Delta I_a} \quad (8-9)$$

for a fixed voltage at the grid ($V_g = \text{constant}$).

The AC anode resistance is expressed in ohms. Therefore, in formula (8-9) ΔV_a and ΔI_a should be expressed in volts and amperes, respectively. For instance, if an increase in voltage $\Delta V_a = 10$ V has caused an increase in anode current $\Delta I_a = 0.5$ mA $= 0.5 \times 10^{-3}$ A, the anode resistance will be $R_i = 20,000$ ohms $= 20$ kilohms.

It is useful to remember the following definition: *the AC anode resistance expressed in kilohms indicates by how many volts the anode voltage should be increased in order to increase the anode current by 1 mA.*

Typical triodes have AC anode resistances from 0.5 to 100 kilohms.

Like the grid-through, the AC anode resistance R_i is not an independent parameter of a triode. If the amplification factor μ and the transconductance S are known, this means that the AC anode resistance can be also determined. In fact, by using a simple algebraic transformation we find that

$$R_i = \frac{\Delta V_a}{\Delta I_a} = \frac{\Delta V_a}{\Delta V_g} \cdot \frac{\Delta V_g}{\Delta I_a} = \frac{\mu}{S}$$

Assume, for instance, that a valve has transconductance $S = 2$ mA/V and amplification factor $\mu = 40$. Hence, its AC anode resistance is

$$R_i = \frac{\mu}{S} = \frac{40}{2 \times 10^{-3}} = 20,000 \text{ ohms}$$

Consequently, the triode has only two independent parameters. Constructional changes affecting the value μ or S also affect the value R_i .

The above relationship between the triode parameters is often given in the following form (Barkhausen formula):

$$\left. \begin{aligned} \mu &= SR_i \\ DSR_i &= 1 \end{aligned} \right\} \quad (8-10)$$

These equations are true, of course, only when all parameters refer to one and the same portion of the characteristic

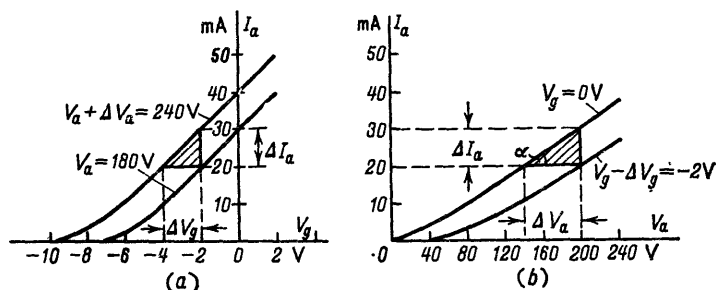


Fig. 8-20. Determining triode parameters

(a) from anode-grid characteristics; (b) from anode characteristics

curve. Generally, these parameters are specified for the steep (approximately linear) portion of the curve.

If we are given with a family of measured grid or anode characteristic curves of a triode, we may easily find its parameters for any operating conditions which are covered by this family of curves.

Consider a family of anode grid characteristic curves shown in Fig. 8-20a. Let the triode parameters have to be calculated for the operating conditions determined approximately by voltages $V_a \approx 210$ V and $V_g \approx -3$ V. From this family of characteristic curves we select the two which cover the region of interest. One of the characteristics corresponds to $V_a = 180$ V, and the other, to $V_a + \Delta V_a = 180 + 60 = 240$ V. These characteristic curves lie in the region of negative grid voltages. To determine the triode parameters, we construct the characteristic triangle such that the operating region of interest appears within this triangle.

The horizontal side of this right-angled triangle (ΔV_g) indicates to what extent the grid voltage should be changed to retain anode current at the same level when transferring

from the characteristic curve for V_a to that for $V_a + \Delta V_a$. The vertical side of the triangle (ΔI_a) may have a two-fold meaning. First, it indicates the increase in anode current when increasing the anode voltage by ΔV_a while maintaining the grid voltage constant (in this case $V_g = -2$ V). Second, this leg also indicates the increase in anode current when increasing the grid voltage by ΔV_g while maintaining the anode voltage constant (in this case $V_a = +240$ V). Thus, we can find all the values required for determining the triode parameters. Finding from Fig. 8-20a the numerical values of the current increments, we can calculate

$$\mu = \frac{\Delta V_a}{\Delta V_g} = \frac{60}{2} = 30$$

$$S = \frac{\Delta I_a}{\Delta V_g} = \frac{10}{2} = 5 \text{ mS}$$

$$R_i = \frac{\Delta V_a}{\Delta I_a} = \frac{60}{10 \times 10^{-3}} = 6000 \text{ ohms (6 kilohms)}$$

Transconductance would decrease and AC anode resistance would increase as we approach the lower knees of the characteristic curves. In catalogues these parameters are usually indicated for the operating conditions under which a given valve should be used.

Let us consider briefly the procedure of determining the triode parameters from a family of anode characteristic curves (Fig. 8-20b). The horizontal side, ΔV_a , of the characteristic triangle indicates to what extent anode voltage should be changed to retain anode current at the same level when transferring from the characteristic curve for V_g to that for $V_g - \Delta V_g$. The vertical side, ΔI_a , has a two-fold meaning. First, it indicates the current increase when the grid voltage is increased by ΔV_g with anode voltage held constant (in this case $V_a = 200$ V). Second, it indicates the current increase when the anode voltage is increased by ΔV_a with the grid voltage held constant (in this case $V_g = 0$ V). Thus, all the values required for determining the triode parameters can be found from the characteristic triangle. The reader is invited to find the numerical values of these parameters by himself from the characteristic curves given in Fig. 8-20b.

It should be emphasized here that the slope of the anode characteristic curve determines the *AC anode resistance* of a valve rather than its *transconductance*. In fact, the less the slope (the less the angle α), the higher the AC anode resistance R_i .

The method of characteristic triangle would be too complicated for use in determining valve parameters during mass production, though the methods of automatic measuring the triode parameters are also based on comparison of voltage and current increments.

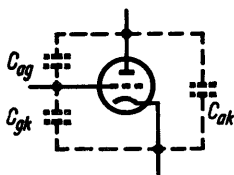


Fig. 8-21. Schematic representation of triode interelectrode capacitances

Apart from the basic triode parameters, μ and S (and auxiliary parameters, R_i and D), one should care of some other parameters. Although being undesirable for certain valve functions, such parameters are intrinsic to any triode because of its geometry. Among these are *interelectrode* capacitances. The anode-cathode (C_{ak}), grid-cathode (C_{gk}) and anode-grid (C_{ag}) capacitances depend upon the sizes of electrodes and the spacing between them. The triode interelectrode capacitances are shown schematically in Fig. 8-21.

For a valve used in an amplifier the capacitance C_{gk} may be called the *input capacitance* since the voltage being amplified is applied to the grid and cathode leads. The capacitance C_{ak} is called the *output capacitance* as the output (amplified) alternating voltage is taken from the anode and cathode leads. Finally, the capacitance C_{ag} should be called the *transfer capacitance* since the input and output circuits may interact through it besides the electron flow. The presence of this signal path may impair the performance of a valve when it is used in a high-frequency amplifier or oscillator. In turn, the capacitances C_{gk} and C_{ak} may change appreciably the design parameters of external (amplifier or oscillator) circuits connected to these electrodes. This is of great importance at high frequencies.

Let us consider the *principle of amplification* with the use of a triode. So that changes in anode current could be proportional to those in grid voltage (i. e., so that a signal could be amplified without non-linear distortions, already referred

to, as applied to Fig. 8-13), the valve should be operated within the linear portion of the anode-grid characteristic curve. Furthermore, in most low-power amplifiers the valve should be operated with no current in its grid circuit. The presence of grid current is associated with useless signal-power consumption in the grid circuit, i. e., with a decrease in amplitude of signal voltage due to the drop of this voltage across the *internal resistance of the signal source*. The role

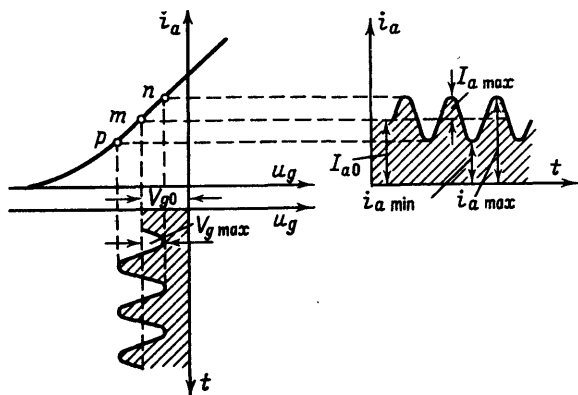


Fig. 8-22. Anode current variations as a function of grid alternating voltage

of grid current will be discussed in more detail later. For a valve to be operated with no grid current, an additional source of DC negative voltage E_g which exceeds all possible signal voltage magnitudes is connected into the grid circuit. This voltage is referred to as *grid bias voltage* or simply *grid bias* (in Fig. 8-22 $E_g = V_{g0}$).

On meeting the specified requirements, the initial point m which determines valve operating conditions in the absence of a signal lies on the linear portion of the characteristic curve (Fig. 8-22). Under these initial conditions the direct current I_{a0} flows in the anode circuit of the valve.

Under the action of a signal alternating voltage having the amplitude $V_{g \max}$ the operating point moves up and down along the characteristic curve between points n and p . This means that the anode current pulsates between $i_{a \min}$ and

$i_{a \max}$, i. e., that an AC component $I_{a \max}$ and the previous DC component I_{a0} flow simultaneously through the anode circuit. In Fig. 8-22 grid voltage and anode current are shown as functions of time.

If valve parameters are known, anode current may be expressed in terms of grid voltage

$$I_{a \max} = S V_{g \max} = \frac{\mu V_{g \max}}{R_i}$$

In compliance with the last equation the action of grid alternating voltage can be represented by a source of alternating EMF $\mu V_{g \max}$ connected into the anode circuit of a valve which has the AC anode resistance R_i . In such a circuit, however, valve amplifying properties cannot be used since there is no load resistor across which the amplified voltage could be produced. In other words, in the case illustrated in Fig. 8-22 only an anode supply battery having EMF E_a is connected between the anode and cathode (see Fig. 8-17). This battery presents no resistance to alternating current and, therefore, for this the anode circuit is found to be shorted.

In order to use valve's amplifying properties, it is necessary to connect into the anode circuit a load resistance (for instance, an ordinary resistor) from which is taken the amplified alternating voltage. When an alternating voltage is applied to the valve input (i. e., between grid and cathode leads) and a load resistor is connected into the anode circuit, the valve is said to be under the *operating condition*.

In the schematic diagram shown in Fig. 8-23 the source of the grid DC voltage E_g and the source of the alternating voltage being amplified $V_{g \max}$ are connected in series into the grid circuit of the amplifying valve. The anode circuit contains the load resistor R_l connected in series with the DC voltage supply E_a . Anode current flows through the load resistor. The direct component I_{a0} of this current produces a voltage drop $I_{a0}R_l$ across the resistor R_l , thus decreasing the anode voltage applied to the valve. Consequently, the DC voltage acting between the anode and cathode leads of the valve is

$$V_{a0} = E_a - I_{a0}R_l$$

Operation of a valve with a load resistor (or any other type of a resistive load) connected in its anode circuit is cha-

racterized by the following fact: the anode voltage decreases as grid voltage is increased and vice versa. In fact, according to the definition of transconductance, formula (8-8), anode current will change by $\Delta I'_a = S\Delta V_g$ when grid voltage is changed by ΔV_g . On the other hand, by the definition of AC anode resistance, formula (8-9), if the anode voltage is changed by ΔV_a , anode current will change by $\Delta I''_a = \Delta V_a / R_i$. Con-

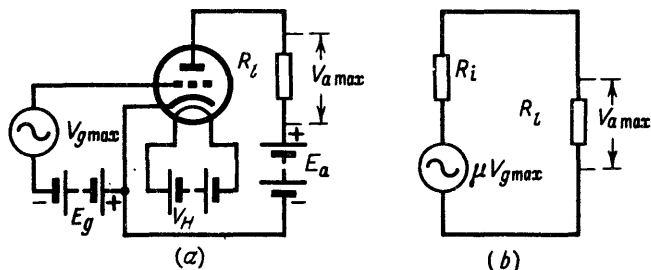


Fig. 8-23. Voltage amplifier built around a triode

(a) simplest circuit; (b) equivalent circuit of amplifier anode circuit for alternating current

sequently, when grid and anode voltages change simultaneously, the total change in current will be

$$\Delta I_a = \Delta I'_a + \Delta I''_a = S\Delta V_g + \frac{\Delta V_a}{R_i} = S \left(\Delta V_g + \frac{\Delta V_a}{SR_i} \right)$$

or taking $\mu = 1/D = SR_i$ from formula (8-10), we find the most important relationship between the anode current and both voltages

$$\Delta I_a = S (\Delta V_g + D\Delta V_a)$$

For the amplifier shown in Fig. 8-23a an increase in grid voltage causes an increase in anode current. As this takes place, the voltage drop across the load resistor increases so that *anode voltage decreases* accordingly. Therefore, the resultant increase in anode current will prove to be less than it would be in the absence of the load resistor ($R_l = 0$) when there would be no reverse effect (reaction) from the anode circuit.

Assume now that a signal alternating voltage is applied to the input of a "loaded" valve. We will consider the ampli-

tude of this voltage as the change in grid voltage, i. e., $\Delta V_g = V_{g \max}$. Similarly, we present the amplitude of anode current by its increase, $\Delta I_a = I_{a \max}$, and the amplitude of alternating voltage at the anode we will call the increase in anode voltage, i. e., $\Delta V_a = V_{a \max}$. Substituting these values in the last formula and taking into account that the voltage across the resistor R_l is subtracted from the anode voltage, we obtain

$$I_{a \max} = S (V_{g \max} - \Delta I_a R_l) = S V_{g \max} - \frac{I_{a \max} R_l}{R_i}$$

By transferring the terms with factor $I_{a \max}$ to the left-hand side of the equation and taking $I_{a \max}$ out of the brackets, we can write

$$I_{a \max} \left(1 + \frac{R_l}{R_i} \right) = S V_{g \max} = \frac{\mu V_{g \max}}{R_i}$$

or by using Ohm's law, we find

$$I_{a \max} = \frac{\mu V_{g \max}}{R_i + R_l} \quad (8-11)$$

Thus, by considering only voltage and current alternating components in the anode circuit, we can construct an equivalent circuit as shown in Fig. 8-23b. In this figure the amplifier is represented by an alternating EMF source having an amplitude $\mu V_{g \max}$ and AC anode resistance R_i . This source operates on the external resistor R_l .

Let us determine how many times the input AC voltage can be amplified by such a circuit.

Voltage at the amplifier output, i. e., across the resistor R_l (amplified signal voltage) is

$$V_{a \max} = I_{a \max} R_l = \frac{\mu V_{g \max}}{R_i + R_l} R_l$$

The ratio of the output alternating voltage to the signal voltage applied to the grid we shall call the *gain* (or *voltage gain*) of the amplifier. It is defined as

$$K = \frac{V_{a \max}}{V_{g \max}} = \mu \frac{R_l}{R_i + R_l} \quad (8-12)$$

Consequently, the voltage gain can be computed provided the valve parameters and load resistance are known. Assume,

for instance, that under the operating conditions chosen the valve has $\mu=25$ and $R_i=12,000$ ohms and the load resistance is $R_l=8000$ ohms, then the gain $K=10$. It is quite evident that for any load resistance R_l the gain K is always less than μ since some part of EMF is dissipated across the anode resistance R_i (see the equivalent circuit in Fig. 8-23b).

In order to determine the rate of change of the anode current in a loaded valve with changing grid voltage, we divide both sides of formula (8-11) by $V_{g \max}$ and call the ratio $\Delta I_a / \Delta V_g$ the slope of a dynamic characteristic curve or the dynamic transconductance,

$$S_D = \frac{\Delta I_a}{\Delta V_g} = \frac{I_{a \max}}{V_{g \max}} = \frac{\mu}{R_i + R_l} = S \frac{R_i}{R_i + R_l} \quad (8-13)$$

Consequently, the dynamic transconductance is less than the static transconductance of the same valve, moreover, it is the less, the greater the load resistor R_l as compared with R_i .

Figure 8-24 illustrates a family of static characteristic curves. The initial conditions are determined by the position of point m on that characteristic curve, which corresponds to the grid bias E_g and anode voltage

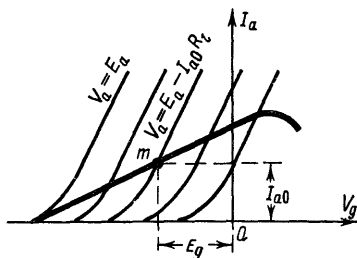


Fig. 8-24. Dynamic characteristic of valve loaded by a resistor

$$V_a = E_a - I_{a0}R_l = V_{a0}$$

As the grid voltage varies to either side from $E_g = V_{g0}$, the anode current changes in compliance with the dynamic characteristic whose slope is determined by formula (8-13) and which is shown in Fig. 8-24 by a thick line. Note that the dynamic characteristic together with the static one corresponding to $V_a = E_a$, where E_a = the total anode battery voltage, originate in the region of negative grid voltages (Fig. 8-24). In the region of positive grid voltages the dynamic characteristic may descend somewhat due to low residual anode voltages (see Fig. 8-18c).

The dynamic characteristic in Fig. 8-24 refers to the valve loaded with the resistor R_l whose resistance is considered

the same both for direct and alternating anode currents. In those cases when the load contains reactances, certain specific features may be found when constructing the dynamic characteristic curve.

Now, for the same stage (Fig. 8-23) containing the load resistor R_l we consider the representation of operating conditions in terms of a family of anode characteristic curves of the triode. Assume that

the anode battery voltage E_a and bias voltage $V_{g0} = E_g$ are known for the valve whose anode characteristic curves are shown in Fig. 8-25. Imagine first that the triode is shorted ($R_l = 0$). Then the triode will have the voltage E_a at its anode as there is no voltage drop in the external portion of the anode circuit. The anode voltage E_a remains unchanged with any changes in grid

voltage and, therefore, all changes in current (displacements of the operating point) will be represented by the vertical line ABC . If the resistor R_l is connected into the circuit, then with no anode current (i. e., when the valve is cut off by the bias) the anode voltage E_a will remain unchanged. If the grid voltage is increased (i. e., negative bias is decreased), the anode current will rise and the voltage drop across the resistor R_l will increase proportionally to it. The anode voltage with respect to the cathode,

$$V_a = E_a - I_a R_l$$

will decrease accordingly and the operating point will move from A along the straight line passing through the point m . This straight line intersects the vertical axis at the point n cutting off the segment E_a/R_l . The straight line An may be called the dynamic characteristic or *load line*. The initial operating conditions are determined by the point m which lies on the characteristic curve corresponding to the initial bias E_g . Under the initial conditions the anode current has

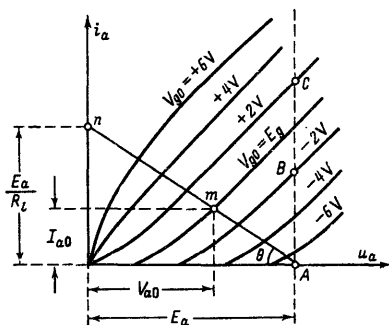


Fig. 8-25. Load line determined from a family of anode characteristics

the value I_{a0} and the anode voltage, the value V_{a0} . The higher the load resistance R_L , the smaller the segment cut off on the vertical axis and, consequently, the smaller the angle θ .

Let us return again to the equivalent circuit shown in Fig. 8-23b. As known from electrical engineering, any active device may be presented either by an EMF source or by a current source. The same refers to active devices used in radio circuits. The EMF source $\mu V_{g \max}$ having the internal resis-

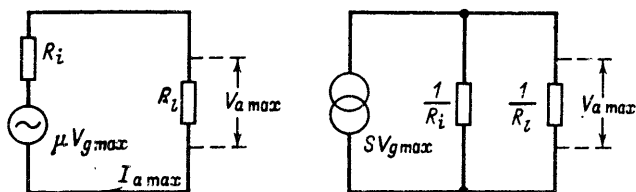


Fig. 8-26. EMF source and equivalent current source

tance R_i may be replaced by the current source $SV_{g \max}$ with the internal conductance $G_i = 1/R_i$ as illustrated in Fig. 8-26. Let us show that the current I_a through the resistor R_L and the voltage $V_{a \max}$ remain unchanged with such replacement.

In the circuit provided with the EMF source the current through the load resistor was

$$I_{a \max} = \frac{\mu V_{g \max}}{R_i + R_L}$$

and had its maximum value with the anode circuit shorted (i. e., when $R_L = 0$)

$$I_{sh} = \frac{\mu V_{g \max}}{R_i} = SV_{g \max} \quad (8.14)$$

In the circuit provided with the current source conductances $1/R_i$ and $1/R_L$ are connected in parallel. Current is distributed proportionally to the conductances so that in the external conductance $(1/R_L)$ it is equal to

$$I_{a \max} = \frac{I_{sh}}{\frac{1}{R_i} + \frac{1}{R_L}} \cdot \frac{1}{R_L} = \frac{I_{sh} R_i}{R_L + R_i} = \frac{\mu V_{g \max}}{R_i + R_L}$$

We see that the current through the load is the same in both cases, and, therefore, the two circuits are equivalent. One should bear in mind, however, that the entire internal conductance of the current source is represented by the conductance $1/R_i$, i. e., the source itself has no internal conductance (in other words, the source has an infinite internal resistance).

Returning to formula (8-12) and considering formula (8-14), the voltage gain can be expressed as follows

$$K = \frac{V_{out}}{V_{in}} = \mu \frac{R_l}{R_i + R_l} = \frac{S}{G_i + G_l} = \frac{S}{G_t} \quad (8-15)$$

where

$$G_i = \frac{1}{R_i}; \quad G_l = \frac{1}{R_l}; \quad G_t = G_i + G_l = \frac{R_i + R_l}{R_i R_l}$$

It is worth while to repeat here the numerical example given earlier for formula (8-12). For a valve with parameters $\mu=25$ and $R_i=12,000$ ohms and a load resistance $R_l=8000$ ohms, we have $S=\mu/R_i=25/12 \times 10^{-3}=25/12$ mS and $G_t=1/8000+1/12,000=5/24$ mS. Hence,

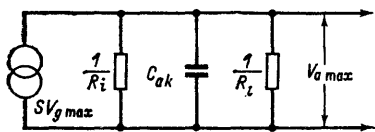


Fig. 8-27. Amplifier equivalent circuit considering valve output capacitance

$$K = \frac{S}{G_t} = \frac{25}{12} : \frac{5}{24} = 10$$

which is identical with the result obtained from formula (8-12).

The physical concept of the EMF source, of course, is more customary than that of the current source. In those cases, however, when a circuit contains some parallel conductances, calculations for the circuit with the current source prove to be much simpler. Let us consider the same resistance-loaded amplifier but operating at such high frequencies that the valve output capacitance C_{ak} should be taken into account. Its equivalent circuit is illustrated in Fig. 8-27. The total conductance in this case proves to be a complex one and equal to

$$\dot{Y}_t = \frac{1}{R_i} + \frac{1}{R_l} + j\omega C_{ak} = G_t + j\omega C_{ak}$$

The processes occurring in the anode circuit of the amplifying triode were discussed above. Let us consider now the processes in the grid circuit of the triode. It is seen from Fig. 8-23 that, apart from the source of signal voltage $V_{g \max}$, the grid is fed with a negative bias E_g which must be always greater than the maximum signal voltage ($|E_g| > V_{g \max}$). Such a bias prevents grid current from appearing in the circuit.

What is the hazard from grid current if no bias is applied to the grid? In Fig. 8-28a* the grid circuit is presented with

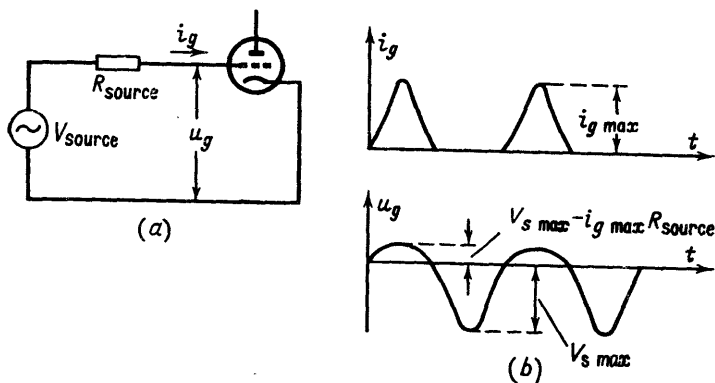


Fig. 8-28. Processes occurring in grid circuit in the absence of bias
(a) grid circuit diagram with signal source; (b) input voltage distorted waveform

no bias but in the presence of a signal source producing the voltage $V_{s \max}$ and having the internal resistance R_{source} . The grid current will appear only during positive voltage half-cycles since the triode grid characteristic, similarly to the diode characteristic, exists only in the region of positive voltages (see, for example, Fig. 8-18). The grid current produces some voltage drop across the source internal resistance R_{source} . There is no grid current during negative half-cycles, so that the entire source voltage is applied to the valve input (i. e., between the grid and cathode). This means

* In Fig. 8-28 and hereinafter the cathode is assumed to be heated (either directly or indirectly) and, therefore, the filament battery is not shown.

that the input voltage V_g will be nonsinusoidal (Fig. 8-18b). Its negative and positive peaks are equal to $V_{s \max}$ and $V_{s \max} - i_{g \max} R_{\text{source}}$, respectively. This distorted grid voltage will distort the waveform of the alternating anode current even if the operating portion of the dynamic characteristic curve is a straight line. From this it follows that the application of a negative bias voltage is indispensable for undistorted operation of the triode.

At the same time the presence of grid current implies the signal power consumption in the amplifier input circuit, i. e.,

$$P_{in} = V_{in}^2 G_{in} = I_{in}^2 R_{in} \quad (8-16)$$

This power is useless for the control of anode current but it loads the signal source. If the negative bias voltage is applied and grid current is zero, then $P_{in} \rightarrow 0$, i. e., power amplification clearly takes place, and there is even no need to evaluate it (power amplification is a compulsory feature of any amplifier). In very high-frequency valves, however, even in the presence of bias, power consumption exists associated with losses in dielectrics and other causes. Therefore, at very high frequencies the power gain must be computed, which is

$$K_p = \frac{P_{out}}{P_{in}}$$

Special very high-frequency triodes will be described in Chapter X.

8-5. MULTIGRID ELECTRON VALVES

Triodes find application in low-frequency amplifiers of receivers, in special very high-frequency amplifiers, in a variety of pulsed circuits and, sometimes, in power stages of transmitters. However, because of its inherent disadvantages the triode is not the principal type of electron valve used in radio equipment.

The first disadvantage of the triode is the substantial anode-grid capacitance (C_{ag}). In any case this capacitance amounts to several picofarads and offers noticeable susceptance ωC_{ag} between the anode and grid circuits at sufficiently high frequencies (Fig. 8-29). The undesired effect of this capacitance is the fact that under the action of alternating voltage appearing at the anode in the presence of a load resis-

tor, an alternating anode current branches off through this capacitance into the grid circuit. This means that a portion of the output voltage returns to the valve input, thus resulting in the so-called *feedback*. These two voltages (signal and feedback voltages) acting simultaneously at the valve input may sharply and adversely affect amplifying properties of the valve.

Another disadvantage of the triode is its low amplification factor μ . It seems that the amplification factor of any desired value can be obtained by using a sufficiently dense grid. In this case, however, even a slight negative voltage applied to the grid will cut off the valve and, therefore, its characteristics (see Fig. 8-19a) will be located, for the most part, to the right of the coordinate origin, i. e., in the region of positive grid voltages. Triodes having a high amplification factor μ are usually called the "right-hand" valves, i. e., the ones which operate as an amplifier only in the *presence of a grid current*.

Both of these disadvantages are eliminated when a second grid called the *screen grid* is placed between the anode and the first (or control) grid. The electron valve with four electrodes is called the *tetrode*. One of the possible constructions of a tetrode and the diagram of its connection into a circuit (with the resistor R_L in the anode circuit) are shown in Fig. 8-30. The screen grid which encloses fully the control grid and cathode is mounted on a horizontal flat metal disk. The function of the screen grid and the disk is to reduce the capacitance between the anode and the first grid and also between their leads and connecting wires.

To explain physically the screening effect of the second grid, let us consider the action of an anode alternating voltage on the control grid circuit. In Fig. 8-31a the grid and anode of the triode are shown as two plates of the capacitor C_{ag1} . The "capacitor" is under the action of an output (amplified) voltage $V_{a \max}$ and allows a reverse alternating current

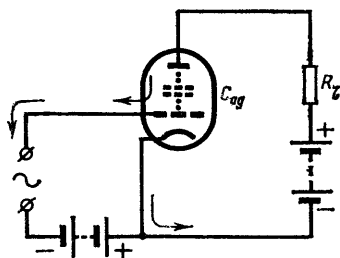


Fig. 8-29. Effect of anode-grid capacitance on triode performance

$I_{rev 1 \max} = V_{a \max} \omega C_{ag1}$ to flow into the grid circuit. In Fig. 8-31b the grid $g2$ is placed between the control grid $g1$ and anode a . Grid $g2$ is made in the form of a metal plate shorted to cathode k . The electric lines of force coming from the anode terminate at the screen so that the reverse current $I_{rev 2 \max} = V_{a \max} \omega C_{ag2}$ is built up in its circuit. If the screen

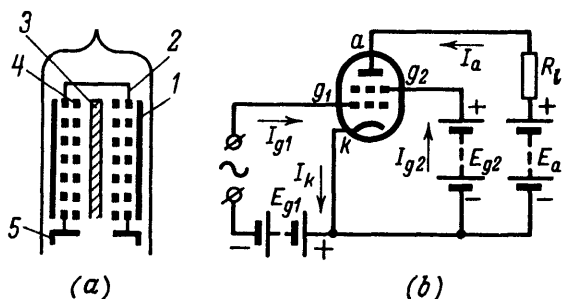


Fig. 8-30. Tetrode

(a) tetrode construction (1—anode; 2—screen grid; 3—cathode; 4—control grid; 5—screen disc); (b) electrode supply circuit

were solid, no electric lines of force from the anode could reach the first grid. No alternating voltage at grid $g2$ with respect to grid $g1$ is present due to the short circuit to the cathode. Under these conditions current $I_{rev 1 \max}$ in the control grid circuit capable of disturbing normal operation

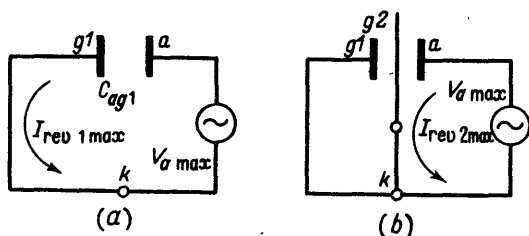


Fig. 8-31. Eliminating the valve transfer (anode-grid) capacitance with the aid of screen

of the amplifier would be equal to zero. In other words, stray capacitance C_{ag1} would be absent.

Such complete screening with a screen-to-cathode short circuit would disturb, however, the principal functions of

the valve: electrons could not flow from the cathode to anode, and the valve could not operate. Therefore, the *screen is a grid* through which electrons can pass easily. As this takes place, some electric lines of force coming from the anode will certainly penetrate through the screen grid reaching the control grid and then the cathode. Proceeding from this, three consequences may be derived.

First, capacitance C_{ag1} still exists in the tetrode but its magnitude is drastically reduced (dozens of times) by the shielding action of the screen. In typical triodes this capacitance amounts to several picofarads, whereas in tetrodes it constitutes only fractions of a picofarad.

Second, tetrodes have a *high amplification factor* μ . This can be easily illustrated numerically. Assume that the screen grid has the grid-through $D_2=1/50$. This means that only a 1/50th portion of the electric lines of force coming from the anode will pass through the screen grid. Then assume that the control grid (usually having a greater spacing between its wires) has the grid-through $D_1=1/8$. In other words, a 1/8th portion of electric lines of force will pass through this grid to the cathode. Consequently, only 1/8th from a 1/50th portion of the electric lines of force coming from the anode will reach the cathode, i. e., a 1/400th portion. Thus, the resultant grid-through for the tetrode is $D=D_1D_2$. The amplification factor can be calculated from the following equation

$$\mu = \frac{1}{D} = \frac{1}{D_1D_2}$$

In our example $\mu=400$. For typical values of transconductance the tetrode has a higher AC (anode) resistance as well. For instance, if transconductance $S=4$ mS, the AC resistance for our example will be

$$R_i = \frac{\mu}{S} = \frac{400}{4 \times 10^{-3}} = 100,000 \text{ ohms}$$

Thus, screen-grid valves feature the increased amplification factors μ and AC resistances R_i as compared with the same parameters of three-electrode valves (provided that the transconductance which depends largely upon the structure of the cathode and control grid is the same for both valves).

Third, the screen grid-to-cathode short circuit proves to be inadmissible. The fact is that when the grid-through D is low, the anode field exerts a very weak effect upon the potential barrier occurring near the cathode. If voltage applied to the control grid is negative, the anode accelerating field will be fully neutralized and *anode current* will prove to be equal to zero when zero voltage is applied to the screen grid. To overcome the potential barrier in the presence

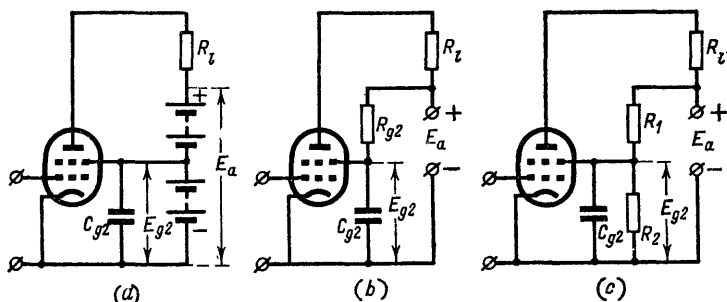


Fig. 8-32. Variants of screen-grid supply

(a) from anode battery; (b) through damping resistor; (c) through voltage divider

of negative first-grid voltage a *DC positive voltage* E_{g2} has to be applied to the screen grid. The voltage applied to the screen grid of a tetrode is usually

$$E_{g2} = (0.25 \text{ to } 0.5) V_{a0}$$

It is this voltage that accelerates the electrons in their path from the cathode to the anode. In Fig. 8-30 *b* voltage E_{g2} is supplied from a separate battery.

It is easy to understand that some electrons are intercepted by the screen grid and current I_{g2} arises in its circuit so that the cathode current I_k turns out to be the sum of three currents. Current I_{g2} serves no useful purpose.

No separate battery is generally used to supply voltage to the screen grid. Voltage E_{g2} is taken from the anode supply source. Fig. 8-32 illustrates possible variants of screen-grid supply from the anode battery: by means of a lead from a required section of the anode battery, by means of a series resistor damping excess battery voltage ($I_{g2}R_{g2}$), and by means of a voltage divider composed of two resistors (R_1

and R_2). The first method presents certain difficulties in practice, the second one is suitable for receivers having battery supply, and the third method along with the second one is applied in receivers powered from AC mains through a rectifier.

Capacitor C_{g2} must be connected between the screen grid and the cathode to short the screen grid to the cathode at high (signal) frequency. It is precisely due to this short circuit that the second grid features properties of the screen which reduces the reverse action of anode voltage upon the first grid circuit in compliance with Fig. 8-31b. What is more, under the action of a signal upon the flow of electrons an alternating component arises both in the anode current and screen-grid current. While passing through the damping resistor R_{g2} , this alternating current builds up a signal frequency alternating voltage across the screen grid (this time by "electronic" way rather than by the capacitive one). Alternating voltage across the screen grid is in antiphase with respect to the voltage across the control grid, thus reducing signal gain. It is the short circuit (through capacitor C_{g2} to the cathode) for alternating current that allows this harmful effect to be eliminated.

The capacitance of C_{g2} should satisfy the inequality

$$\frac{1}{\omega C_{g2}} \ll R_{g2}$$

The capacitance of C_{g2} amounts, for all practical purposes, to thousands or dozens of thousands of picofarads for radio-frequency signals and to several microfarads for audio-frequency signals. As far as the damping resistor R_{g2} is concerned, its resistance can be determined from the equation

$$R_{g2} = \frac{E_a - E_{g2}}{I_{g2}}$$

where I_{g2} is the screen-current direct component. In low-power valves resistors R_{g2} have resistances of not less than hundreds of kilohms.

The second disadvantage of a triode (the "right-hand" location of characteristics when the amplification factor μ is high) is also eliminated in the tetrode due to the presence of the screen grid. As stated above, the total amplification

factor of a tetrode proves to be higher owing to a low grid-through for both tetrode grids. Moreover, the screen grid-through with respect to the cathode (through the control grid) is much greater than for the anode, and, therefore, the flow of electrons in the valve is determined not so much by anode voltage as by the DC voltage applied to the screen grid. This grid serves as an "anode" through which a greater portion of electrons pass and hit the principal anode.

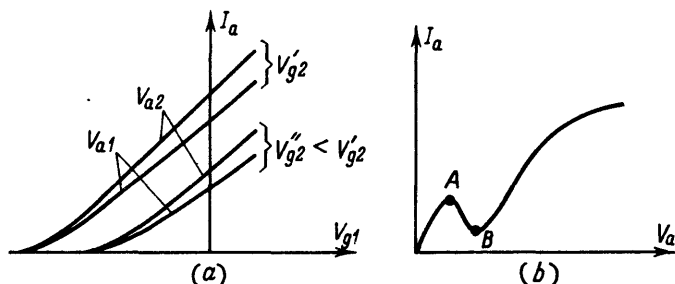


Fig. 8-33. Characteristic curves of tetrode

(a) relationship between anode current and control grid voltage; (b) relationship between anode current and anode voltage

In Fig. 8-33a two sets of characteristic curves are shown for two values of the voltage V_{g2} but for the same anode voltage (for both sets). The two sets are shifted appreciably one from the other. Each set consists of two characteristic curves taken at different values of the voltage V_a . As seen from Fig. 8-33, changes in the voltage V_a affect negligibly the characteristic curves causing them to diverge slightly one from the other. This fact indicates that the valve has a high amplification factor. At the same time, by properly selecting the voltage V_{g2} the family of anode-control grid characteristic curves can be shifted sufficiently to the left for normal values of the anode voltage V_a .

An ordinary tetrode has also its own disadvantage which prevents wide use of this type of valve. If anode voltage becomes less than screen-grid voltage during amplifier operation, a sharp drop in anode current is generally observed. This phenomenon is known as the *dynatron effect* graphically illustrated in Fig. 8-33b on the anode characteristic curve. This characteristic curve shows a minimum and two bends.

On the descending portion *AB* of this curve an increase of anode voltage results in a decrease of anode current, thus causing distortion of the signal being amplified.

The cause of the dynatron effect lies in secondary emission from the surface of metals. When sufficiently fast electrons strike the metal surface, some energy of electrons is spent on heating the metal, but some of their energy is transferred directly to free electrons inside the metal giving them the possibility to escape into the external space. This effect is called the *secondary emission*. One fast electron can knock several secondary electrons out of the metal.

Secondary emission may also occur from the anode surface in a triode, but here it is not noticeable since slow secondary electrons cannot move away from the anode as it attracts the emitted electrons back (as many as electrons are emitted from the anode, the same number returns back). As far as the tetrode is concerned (see Fig. 8-33*b*), when anode voltage rises from zero, anode current first increases since primary low-speed electrons cannot knock out secondary ones. Secondary emission arises with the increasing anode voltage. The number of dislodged electrons exceeds that of the primary electrons, thus anode current decreases. Secondary electrons are attracted by the screen grid whose potential is higher than that of the anode, and the current flowing through the screen-grid circuit increases. With the ever increasing anode voltage, a greater portion of secondary electrons return back to the anode so that the anode current increases. Furthermore, with substantially high voltages applied to the anode, anode current rises additionally at the expense of the attraction of the secondary electrons being knocked out of the screen grid.

How can the dynatron effect be eliminated in the tetrode?

One of the possible ways is a beam-power construction or simply a *beam tetrode*. In such a tetrode both grids have the same number of turns arranged precisely in line (against each other). Owing to such a construction, the electrons are formed into flat beams after being focused by the first grid and pass through the second grid without striking its wires, thus decreasing the amount of useless current in the screen-grid circuit. In the beam-power tetrode the anode has the form of a cylinder with a large radius, and its internal surface is

at a distance from the flat sides of the screen grid. As a consequence, the force of attraction exerted by the screen grid on the secondary electrons becomes weaker. To direct electrons only through flat sides of the grid, i. e., over the longest path to the anode, the beam-power tetrode is provided with vertical metal plates S connected to the cathode, i. e., having zero potential. These plates repel electrons, thus assisting in producing the "fan-shaped" beams (Fig. 8-34a).

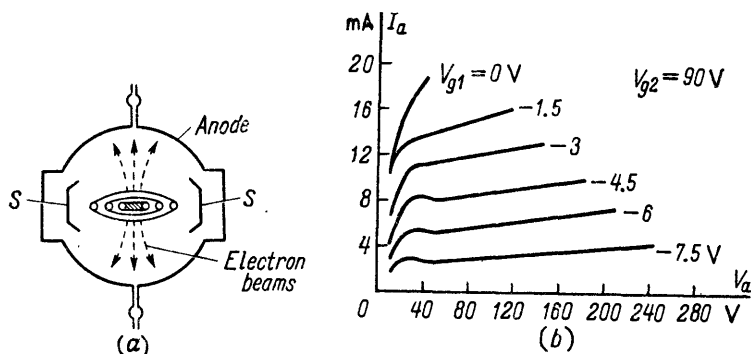


Fig. 8-34. Beam tetrode

(a) tetrode construction and operating principle; (b) family of anode characteristic curves

If in the beam-power tetrode voltage at the screen grid is higher than that at the anode, primary electrons are slowed down in the space between the anode and the screen grid, thus forming a negative space charge. The latter repels the secondary electrons dislodged from the anode back to the anode. This means, that the space charge functions as a protective shield. In an ordinary tetrode the flow of electrons is dispersed by grid wires and cannot produce such a space charge.

Fig. 8-34b illustrates anode characteristic curves of a beam-power tetrode. Owing to the construction described above, the minima caused by the dynatron effect are quite negligible. Flat portions of characteristic curves originating for this valve at small anode voltages serve as its operating portions. Beam-power tetrodes are generally used in the output stages of receivers and transmitters, including those of sufficiently high power.

Another method used for eliminating the dynatron effect consists in the following. An additional (third) grid is placed on the path of electrons between the screen grid and the anode. Being electrically connected to the cathode, this grid has zero potential which prevents the flow of secondary electrons from the anode to the screen grid. The additional grid is called the *suppressor grid* or simply *suppressor*. An electron valve with five electrodes is called the *pentode*.

Advantages of the pentode are its low transfer (anode-control grid) capacitance, left-hand location of anode-control grid characteristics at low anode voltage, and the elimination of the dynatron effect; these facts allow the pentode to be widely used in radio equipment. The schematic diagram of the simplest amplifier built around a pentode is shown in Fig. 8-35.

The amplification factor of pentodes may be as high as 1000 owing to the presence of three grids, and, consequently, anode resistance is usually in the order of about one million ohms. Transconductance is about the same as that for triodes or tetrodes (from 2 to 50 mS for receiver valves). In pentodes transfer capacitance is of the same order as that for tetrodes or less.

High values of anode resistance in pentodes indicate that an increase in anode voltage shifts but slightly the anode-control grid characteristic curves. At the same time when computing the gain for the valve loaded with a resistor from formula (8-12), one may neglect the value of R_i appearing in the denominator as it is very small compared to the very high value of R_i . As a result, we obtain

$$K \approx \mu \frac{R_i}{R_i} = SR_i \quad (8-17)$$

Transconductance S is the most important parameter of a *pentode* when characterizing its amplification properties.

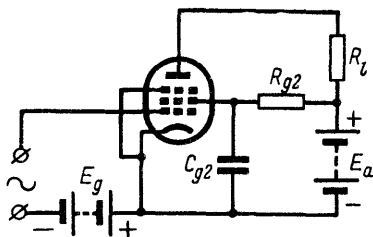


Fig. 8-35. Simplest amplifier employing pentode

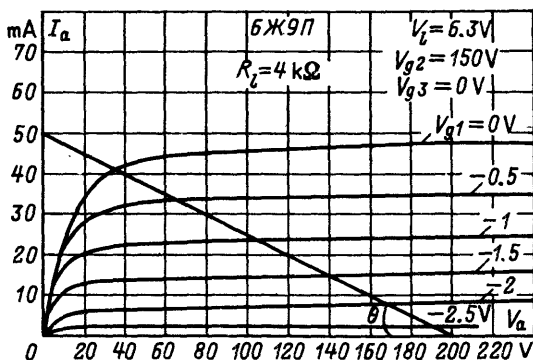


Fig. 8-36. Anode characteristic curves of 6Ж9П valve

Needless to say that the gain of a loaded valve is much less than its amplification factor ($K \ll \mu$).

It can be concluded from the above that dynamic transconductance defined by formula (8-13) will not practically differ from static transconductance. Therefore, it is useless (because of great inaccuracy) to give a graphical representation of the performance of a loaded pentode as was given for a triode (see Fig. 8-24). A more accurate and illustrative representation of the performance of a loaded pentode can be given in terms of a family of anode characteristic curves of a pentode.

In Fig. 8-36 an example of the family of anode characteristic curves of a 6Ж9П type pentode ($S=18$ mS, $R_i=150$ kΩ, $\mu=2700$) is given. The dynamic or load line for the case of the load resistance $R_L=4000$ Ω and the battery voltage $E_a=200$ V is plotted in the same figure.

It will be recalled once again that the graphical representation of valve performance with the aid of a load line may be sometimes more convenient than that expressed in terms of anode-control grid characteristic curves.

Many small signal amplifiers require continuous adjustment of their gain K . As can be seen from formula (8-17), theoretically this aim can be attained by changing smoothly the slope S of that portion of the anode-control grid characteristic curve where the operating point is located. In practice this objective is accomplished by constructing a control

grid in the form of a variable-pitch coil. With such a grid, the negative bias voltage sufficient for cutting off the flow of electrons through closely spaced turns proves to be insufficient to cut off the flow of electrons through widely spaced turns. As a result, the characteristic curve has an elongated "tail" portion (Fig. 8-37) to which the operating point should be shifted when receiving strong signals. In receivers this shift is carried out, as a rule, automatically.

The main efforts directed towards improving receiving amplifier valves refer first of all to the pentodes which are the most widely used type of these valves. Improvements are in progress primarily to increase the transconductance S , mechanical strength (for instance, resistance to vibration) of pentodes and on perfecting other valve characteristics, including service life.

Replacement of grids wound in the form of a coil with *frame grids* made of fine wires stretched tightly over a frame made it possible for the distance between the cathode and the grid to be reduced, thus increasing valve transconductance (for instance, up to 30 mS in the 6Ж23П type pentodes). In another method of increasing transconductance an *additional (fourth) grid* is placed between the cathode and the control grid. A small positive voltage is applied to such a cathode grid with respect to the cathode. As a consequence, the electron cloud is displaced closer to the control (second) grid so that transconductance is increased. For instance, the 6Ж22П type pentode provided with a cathode grid has transconductance of 23 mS. A third method of obtaining high transconductance is the use of *secondary emission* which up to now has been considered as a useless (or even harmful) phenomenon. For this purpose, a *dynode* (secondary emission electrode) is mounted inside the envelope of a tetrode or pentode. The dynode is held positive but at a lower potential than the anode. The primary electrons emitted by the cathode after passing through all the grids strike the specially coated dynode sur-

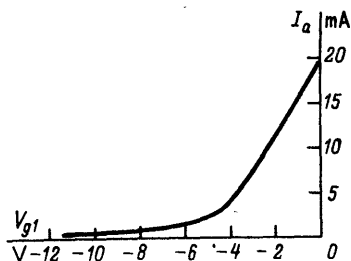


Fig. 8-37. Pentode with tailed characteristic curve

face and knock out from it secondary electrons whose number is greater than that of the primary electrons. Secondary electrons flow towards the anode increasing its current and, consequently, valve transconductance (up to about 30 mS in the 6B11 type valve).

It should be readily apparent that for an amplifier with a given gain the required number of high transconductance valves will be less than that of low transconductance valves.

It is common practice to use *multiunit valves* to reduce the number of valves in radio equipment and simplify circuit wiring. Two (or more) electrode sets are enclosed inside the envelope of such a valve. Among these are *double diodes*, *double triodes*, *diode-triodes*, *diode-pentodes*, and *triode-pentodes*. Each set of electrodes in such valves has appropriate characteristics and performs its own functions.

In summary it should be mentioned that special *valves designed for frequency conversion* and used in superheterodyne receivers are also available. Among these are valves with dual control of electron flow. In other words, these valves have two control grids between the cathode and the anode. A high-frequency signal is applied to one of the control grids while the other grid is connected to a local oscillator. As a result of such a combined control of the electron flow, an oscillation whose frequency is equal to the difference between the signal and local oscillator frequencies can be extracted in the anode circuit. This is basically the process of frequency conversion. A separate valve (usually a triode) is generally employed for the local oscillator. However, a local oscillator valve and a dual control valve may be combined inside a single envelope.

A *heptode* is an example of the dual control valve. The heptode is a seven-electrode valve (such as 6A21 type). It has five grids: two control, two screen and one suppressor. If a triode (intended for use as a local oscillator) is enclosed in the same envelope with the heptode, this valve is called the *triode-heptode* (for instance, 6H11 type).

The processes of frequency conversion and performance of frequency-converter valves will be described in later chapters.

8-6. ELECTRIC CONDUCTION IN SEMICONDUCTORS

Semiconductor materials exhibit remarkable properties making it possible to devise a great variety of semiconductor devices capable of performing important functions in rectifying, amplifying, photoelectric, thermoelectric equipment and in many other fields of engineering.

Semiconductor devices such as *crystal diodes* and *transistors* which perform the same functions as electron valves in low- and medium-power receiving and amplifying equipment are of particular interest to radio amateurs. Among the principal advantages of semiconductor devices are the absence of a heated cathode (i. e., no power for the filament), small size and low weight, mechanical ruggedness, long service life (thousands of hours), low supply voltages (as a rule, 6 to 12 V for receivers) and high efficiency. At the same time, they suffer from a number of disadvantages. Among these are a comparatively low input resistance, spread in parameters between devices of the same type, dependence on temperature, deterioration in performance with the increasing frequency and a number of other properties which are worse than those of electron valves. It should be noted, however, that semiconductor devices are being continually improved.

Semiconductor devices find an ever extending application in radio broadcast receiving equipment and TV sets (excluding picture tubes) and in many other fields of electronics.

In the chapter on electron valves we started from current in a vacuum, now we must discuss the flow of current in a solid substance to impart an understanding of physical processes in semiconductor devices.

Solid substances are classified (roughly) as conductors, dielectrics (insulators) and semiconductors. The essence of this classification consists in the possible ways of transferring electrons of a valence band to a conduction band. Electrons of an atom revolving around the nucleus may have strictly definite (discrete) energy levels although the number of energy levels is unlimited. If no external energy is given to an atom (i. e., it is at 0°K and is exposed to no external radiation), its electrons have minimum energy levels. Electron energy levels W_1 , W_2 , W_3 , etc. are schematically shown in Fig. 8-38a. The outermost (valence) electrons possess

maximum energy. The band encompassing the energy levels of valence electrons is called the *valence band*.

On being excited externally (for instance, by heating the substance), the electron can jump to a higher energy level and, after leaving the given atom, roam at random between atoms. Under the action of an external electric field these free electrons set up an electric current. The energy levels of such electrons comprise a "conduction band" (as shown by the dashed lines in Fig. 8-38a). In conducting materials (me-

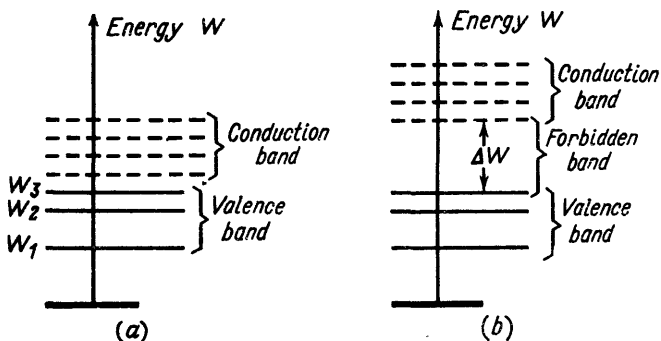


Fig. 8-38. Electron energy levels
(a) in metals; (b) in dielectrics and semiconductors

tals) the *conduction and valence bands lie next to each other*, and practically at all times, even at low temperatures, metals have a great number of electrons imparting the *electric conduction property* to the material.

In *dielectrics* the valence band is separated from the conduction band by the so-called *forbidden band* or energy gap corresponding to energy levels at which electrons cannot exist. The width of the forbidden band ΔW is at least several electron-volts. Only after overcoming this energy gap may an electron participate in conduction current, i. e., only after having acquired external energy greater than ΔW . In dielectrics such electrons are few, so that *their conductivity is negligible*.

In the solids defined as *semiconductors*, energy structure is also characterized by the presence of an *energy gap*. The width (ΔW) of this gap, however, is less than that in dielect-

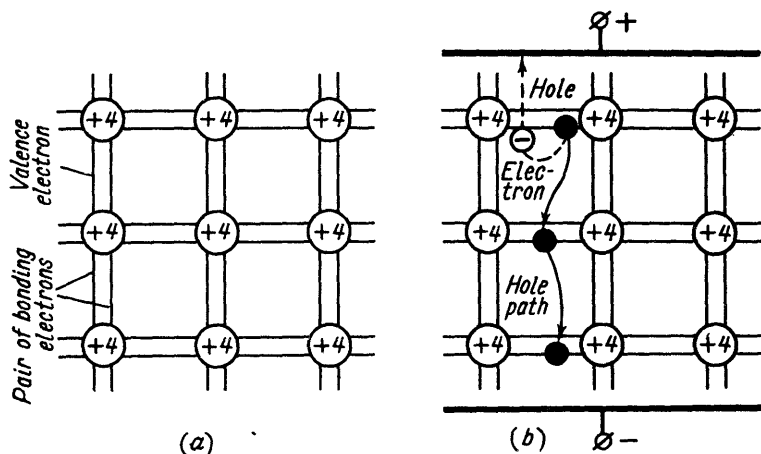


Fig. 8-39. Germanium crystal lattice (a) and hole current in it (b)

rics and does not exceed one electron-volt. Even at room temperature (about 300°K) some valence electrons may jump into the conduction band, thus producing electron current in the semiconductor. Electrons which have lost their energy (for instance, due to cooling) may return back into the valence band, i. e., these electrons do not take part in conduction current through the material. Thus, conductivity of a semiconductor decreases (its resistivity rises) with decreasing temperature; by this property semiconductors differ from metals.

Materials most frequently used in semiconductor technology are germanium (Ge) and silicon (Si) although other semiconductor materials may be applied.

Germanium is an element in Group IV of Mendeleev's periodic system. Consequently, it has four valence electrons in an atom. The remaining 28 electrons in a germanium atom are bonded tightly to the nucleus. Germanium and silicon have a crystal structure: atoms are located at the corners of regular geometrical figures and form a crystal lattice. The crystal lattice of germanium is schematically shown in Fig. 8-39a. Atoms are presented by circles and marked with the valence number (+4). Valence electrons are illustrated in the form of bonding lines with adjacent atoms. Each two

atoms are bonded by a pair of "covalent" (bonding) electrons—one electron from each atom. The covalent bond may be visualized physically in this case as a rotation of electrons in outer shells common for two atoms. The lattice structure under consideration (with no bonding electrons lost) refers to pure germanium (not containing impurities) at 0°K.

Electron bonds in a crystal may be disrupted. One of the causes of this disruption is the heating of a semiconductor which increases oscillations of the atoms near the equilibrium position in the crystal lattice. Having acquired kinetic energy from an external source, an electron may overcome interatomic bonding forces and begin to move at random inside the crystal lattice. This electron may be called the *conduction (band) electron* (as distinct from a free electron in a vacuum). If an electric field is applied across the crystal, conduction electrons tend to move in opposite to the electric lines of force, thus setting up a current through the semiconductor and supply source. This is intrinsic conduction associated with movement of electrons through a semiconductor.

When a valence bond is disrupted and the electron escapes, the unfilled orbit is left in the atom and the atom itself acquires a *positive charge* equal in absolute value to the electron charge. The positive charge produced due to the loss of one electron is called a *hole*. Such an atom state is not stable since an electron from an adjoining atom may jump into this hole and the hole will become filled, i. e., the positive charge will be neutralized. As this takes place, the electron which has escaped from the adjoining atom leaves behind another hole. As a result, holes are capable of moving as positive charges. The electrons filling the holes we shall call *substitution electrons*. Under the action of an external electric field the holes will move (as positive charges) in the direction of lines of force, i. e., from the positive to the negative terminal of the supply source. The motion of the holes constitutes a *hole current*. In Fig. 8-39b the hole motion is shown in the form of disruptions of electron bonds between atoms.

Thus, the current flowing through a semiconductor (as measured by an instrument in an external circuit) will be the *sum of electron and hole currents*. This fact explains the main difference between the current flowing through a semiconductor and the electron current in a vacuum and, consequently,

the difference between semiconductor devices and electron valves. However two circumstances should be borne in mind. First, *both currents* in a semiconductor are essentially *electron currents* as hole current is the motion of substitution electrons in the same direction as that of the conduction electrons. Second, *holes* have *lesser mobility* as they move from atom to atom rather than directly between the source terminals. As a result, in the total current flowing through a semiconductor the hole current constitutes a lesser portion as compared to the electron current.

The number of electrons in a pure semiconductor is equal to that of holes. In a semiconductor *charge carriers*, i. e., electrons and holes, appear in pairs, and the process is known as the generation of electron-hole pairs. Pair generation is inevitably accompanied by recombination of carrier pairs into neutral atoms. Consequently, the concentration (density) of charge carriers, i. e., their quantity per 1 cm^3 , is the same for electrons and holes

$$n_i = p_i$$

where n = concentration of electrons (electron density), i. e., negative charge carriers

p = concentration of holes (hole density), i. e., positive charge carriers. Subscript i denotes that we consider an intrinsic (pure) semiconductor material. For germanium at room temperature $n_i = p_i = 10^{13}$ per cubic centimetre

These are the properties of intrinsic conduction of pure semiconductors. In general, their conductivity is low. To increase conductivity impurities are added to a pure semiconductor. In the crystal lattice impurity atoms replace some portion (small) of the atoms of the parent material ("single crystal"), thus changing significantly its properties.

If a pentavalent element (such as arsenic, antimony, phosphorus) is added to a pure semiconductor as an impurity (a process known as doping), four electrons of each impurity atom form electron-pair bonds with four atoms of the quadrivalent germanium (or silicon), whereas the fifth electron will prove to be an excess one (Fig. 8-40a). This electron can easily detach itself from its atom to become a conduction electron. As a result, the impurity atom becomes a positive ion. This

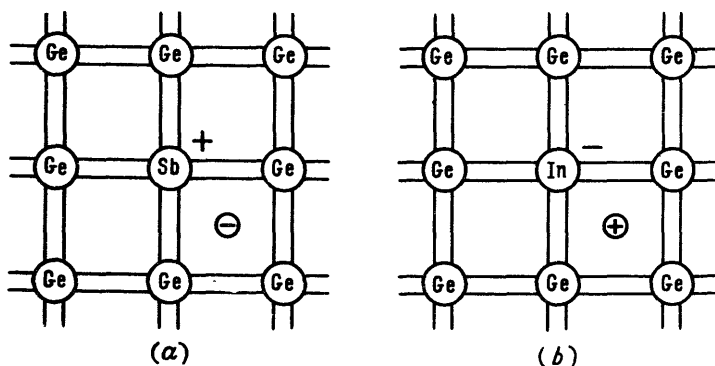


Fig. 8-40. Germanium (Ge) crystals with impurities
 (a) donor impurity (antimony Sb); (b) acceptor impurity (indium In)

means that the pentavalent impurity increases drastically the electron density n . Therefore, this impurity is called a *donor* impurity or simply a donor (since it donates free electrons) and the semiconductor itself becomes the *electronic* or *n*-type semiconductor. In the semiconductor of this type electrons and holes are the *majority* and *minority carriers*, respectively. The hole density p decreases as many times as the electrons density n increases, so that

$$np = n_i^2 = p_i^2 \quad (8-18)$$

A decrease in the number of holes can be explained by the fact that when the number of electrons increases they begin to recombine more frequently with holes which are not replenished.

If germanium (or silicon) is doped with a trivalent element (such as indium, aluminium, gallium, boron), the properties of the single crystal change. Three valence electrons of the impurity atom form electron-pair bonds with three atoms of germanium. The deficient electron for the fourth bond is taken from a germanium atom one of whose valence bonds breaks down and fills all the valence bonds of the impurity atom. As a result, the impurity atom becomes an anion (a negative ion), i. e., a hole is formed in the germanium (Fig. 8-40b). This means that the trivalent impurity increases

the hole density p and, therefore, it is called an *acceptor* impurity or simply an acceptor (since it accepts electrons). The semiconductor itself becomes the *hole* or *p-type semiconductor*. In the *p-type* semiconductor holes and electrons serve as *majority* and *minority carriers*, respectively. The electron density n decreases as many times as the hole density p increases so that formula (8-18) remains valid. It should be noted that the number of impurity atoms should be millions of times less than that of the parent atoms. The production of high-grade pure semiconductors and the accurate proportioning of added impurities constitute a serious technological problem.

When an external electric field is applied across an *n-type* semiconductor bar, electrons start to move towards the positive terminal, thus producing a current in the external circuit. The current of minority carriers (holes) will also take place in the semiconductor but its value is so small as compared to the electron current that it may be neglected. In the bar of a *p-type* semiconductor holes move towards the negative terminal where they are filled with electrons. In this case, the direction of current in the external circuit remains the same as for the *n-type* semiconductor. Here, the current of minority carriers (electrons) may also be neglected.

8-7. P-N JUNCTION AND CRYSTAL DIODES

Two regions, electron and hole, can be obtained by doping a bar of a single-crystal semiconductor with appropriate impurities. The interface between these two regions is known as an *electron-hole junction*, abbreviated *P-N junction*. Such junctions are the basis of any kind of a semiconductor device.

Let us discuss the electron state of the *P-N junction* with no external voltage applied. Assume an abrupt boundary between the regions of electron (*N*) and hole (*P*) conduction (Fig. 8-41a) in the single crystal. Now suppose (although this assumption is not necessary at all) that the electron density in the left-hand region (with donor impurities) and the hole density in the right-hand region (with acceptor impurities) are the same (i. e., $N_d = N_a$, Fig. 8-41b and c). On the opposite sides of the junction there is an excess of carriers of the *N*- and *P*-type, respectively. The electrons

receive thermal energy and diffuse from the N -region to the P -region, while the holes do the same in the opposite direction. This means that majority carriers move away from the layer next to the junction, leaving near it stationary impurity ions which will be positive in the N -region and negative in the P -region. Thus, a double layer of charges which are

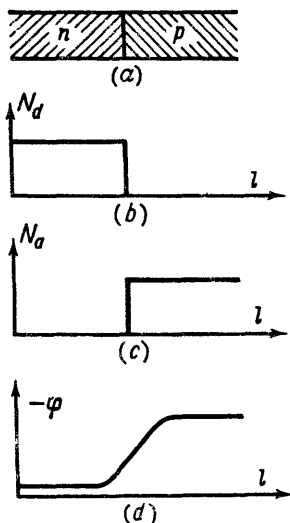


Fig. 8-41. Junction potential barrier

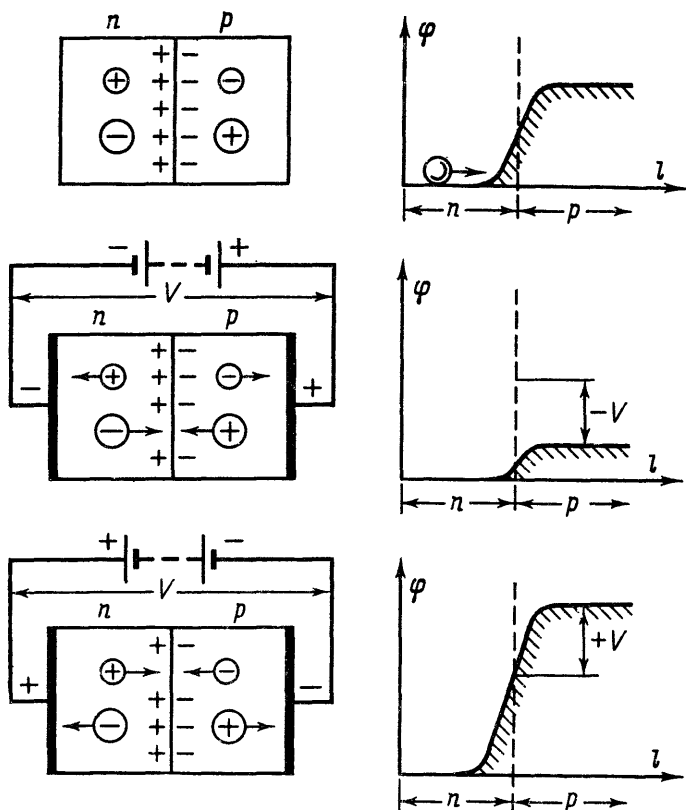
opposite in sign is formed near the contact. The potential difference between the two layers may be called a contact or junction potential difference which sets up an electric field at the junction (Fig. 8-41d) serving as a potential barrier (fractions of volt) to further diffusion of carriers.

With a steady-state potential difference the flow of electrons will not stop completely as the electrons with sufficiently high energy are capable of overcoming the potential barrier. However, under the influence of the electric field across the contact, free electrons available in small quantities in the P -region may drift easily from the right- to left-hand region. The number of these electrons is equal to that of the electrons which move from left to

right through the barrier, i. e., diffusion current is compensated by drift current. This is the equilibrium state.

We have discussed so far the motion of electrons. The same may be said about holes which are available in excess in the P -region and diffuse to the N -region. Holes are charged positively and the electric field set up due to their diffusion from right to left has the same direction as that produced by the diffusion of electrons from left to right. In equilibrium the hole diffusion current is balanced by the hole drift current which has the opposite direction.

There exists some deficiency of electrons and holes on the left and right sides of the junction, respectively. This means that the concentration of carriers at the junction is low (dep-

Fig. 8-42. Current paths in P - N junction

letion layer) and semiconductor resistance here is high ("barrier" region).

The equilibrium state of the P - N junction and the corresponding potential barrier are illustrated in Fig. 8-42 (upper row). Majority carriers are enclosed in large circles, minority carriers are shown by small circles, and ions are not enclosed in circles.

Suppose now that two metal electrodes, whose contact with the semiconductor surface exhibits very low resistance, are attached to the end faces of the semiconductor bar. The po-

sitive and negative terminals of an external battery (Fig. 8-42, middle row) are connected to the electrodes of the P -type and N -type regions, respectively. Under these conditions, current starts flowing through the P - N junction, the connecting wires and the external circuit. From the figure it is clear that the external electric field is opposite to the field produced in the boundary region. In other words, the junction potential difference will be somewhat compensated by the external electric field.

The majority carriers in each semiconductor region (large circles), moving towards the junction under the action of the external field, may overcome the remaining potential difference (the potential barrier) and pass through it. The number of majority carriers passing through the boundary region and, consequently, the value of current in the circuit will increase with the increasing voltage.

If electrode polarity is reversed by interchanging the position of the supply source terminals (Fig. 8-42, lower row), minority carriers (small circles) then tend to move towards the junction from both sides. These carriers may pass through the boundary region even with no external field being applied since the potential difference in this case is "favourable" to their passage. The number of minority carriers, however, is small and, therefore, the current they constitute is negligible. As far as the majority carriers are concerned, the external field tends to enhance the potential barrier.

Thus, the P - N junction resistance depends upon the current direction, namely, with the polarity aiding the majority carriers to pass through the junction, the resistance is low and decreases with the rising voltage; when polarity is reversed, the resistance is high and does not depend practically upon the voltage applied. This property of the P - N junction made it possible to devise a great variety of semiconductor devices, first in the list were rectifiers. Rectifiers using crystals of germanium and silicon as a base are referred to as *crystal diodes*.

Germanium and particularly silicon diodes find extensive application in the AC power rectifiers which feed radio equipment. In their small size and high efficiency such rectifiers have advantages over kenotron type rectifiers and even old types of semiconductor rectifiers (such as copper-oxide and

selenium rectifiers). The so-called junction (alloyed) diode is the main type of a germanium rectifier.

The *construction* of a germanium junction *diode* is shown schematically in Fig. 8-43a. It consists primarily of a germanium crystal which has electron conduction (*N*-type germanium) owing to antimony or arsenic impurities added to it. A tin electrode (holder) is alloyed with the semiconductor on one side of the crystal. This electrode provides tight con-

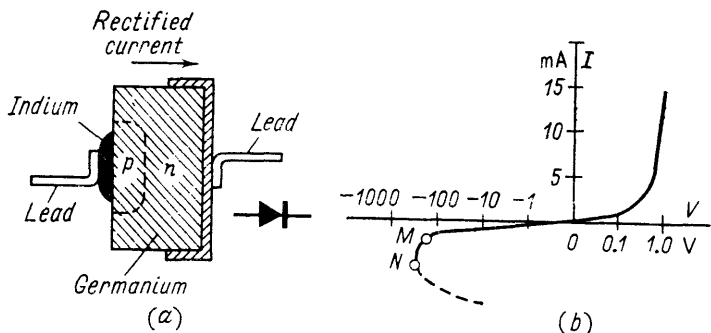


Fig. 8-43. Germanium junction diode

(a) construction and circuit symbol; (b) static characteristic curve

tact and permits an external circuit terminal to be connected. On the opposite side of the crystal an indium bar is alloyed with the semiconductor at a temperature of about 500°C. Indium atoms diffuse (penetrate) into the semiconductor crystal, thus forming in it, after cooling, a hole conduction layer (*P*-type germanium). Thus, a rectifying *P-N* junction is formed between the surface layer of the *P*-type germanium and the remainder of the *N*-type germanium. The second terminal of the external circuit is soldered to the indium bar. The resultant *P-N* junction is referred to as a junction diode because of the relatively large area of the *P-N* junction while the term "alloyed junction" originates from the name of the process of alloying indium with germanium. A diode symbol used in circuit diagrams is given in the same figure.

A static volt-ampere *characteristic* curve of a germanium junction *diode* is shown in Fig. 8.43b. (For clarity, the vol-

tage scale is made nonuniform.) While considering this characteristic, it might be well to point out the following properties of the germanium diode. In the region of positive voltages (i. e., in the "forward" direction) current appears at low voltages (about 0.1 V), and then rises abruptly. Consequently, the slope of the characteristic curve is great and it exceeds that of vacuum diodes. In other words, the forward resistance of a crystal diode is negligible (including the rectified current resistance in the positive region) and, therefore, rectifier efficiency is high (up to 97-99 per cent). In the region of negative voltages, as distinct from vacuum diodes, the crystal diode still conducts a small current (due to minority carriers).

Thus, the characteristic curve of the crystal diode is, as a whole, non-linear since its resistance in the negative region is much higher than that in the positive region.

Under high reverse voltage (about 100 V for a diode as shown in Fig. 8-43), the barrier layer of the *P-N* junction breaks down and reverse current increases. The layer resistance becomes so small that with increasing current the voltage drop across this resistance remains constant and then starts to decrease (the dashed line descending portion of the characteristic curve). When the diode is in operation, reverse current should not exceed values at which the *P-N* junction is overheated. An increase of current in the portion *MN* does not yet damage the diode.

On the basis of the properties described above the following operating parameters of a junction diode can be specified:

(1) the maximum rectified current not causing diode damage (due to overheating) during long-term operation; (2) the peak reverse voltage not causing a non-recoverable breakdown during long-term operation; (3) the DC voltage drop across the diode at the maximum rectified current (diodes with low voltage drops have great advantages); (4) the reverse current at the peak reverse voltage (it is advisable to have values of reverse current as low as possible so that good rectifying properties can be obtained).

Junction diodes are also manufactured on the base of semiconductor crystals of silicon. Silicon diodes are used successfully in a wider operating temperature range than germanium diodes.

Apart from rectifier parameters, a crystal diode is characterized by its capacitance which may be considered as the capacitance of a parallel-plate capacitor in which the depletion layer serves as a dielectric and the plates are formed by current-carrying boundary layers of electron and hole regions. When reverse voltage is increased, the depletion (barrier) layer spreads out and diode capacitance decreases (as if the

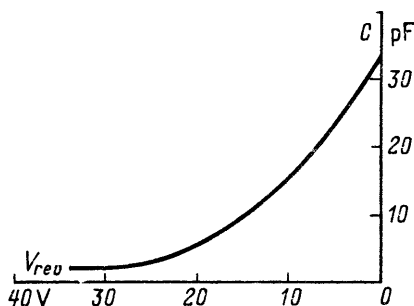


Fig. 8-44. Diode capacitance as a function of reverse voltage

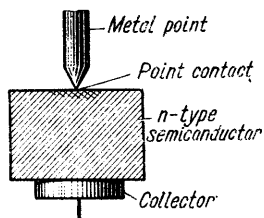


Fig. 8-45. Point-contact diode

plates are spaced more apart). Certainly, diode capacitance is undesirable in principle when the diode is operated as a rectifier since it permits alternating current to flow to the load. However, the capacitance dependence upon the voltage applied (Fig. 8-44) is of great practical importance in a number of applications. For instance, when connected into an oscillatory circuit such a diode having variable capacitance (which is called the *varicap*) enables the circuit to be continuously retuned electrically without using a capacitor having mechanically rotated plates. Such a diode is one of the most important components of a modern parametric amplifier.

The so-called *point-contact diodes* representing diodes with low capacitance are in wide use as non-linear (rectifying) elements for RF-signal detection and frequency conversion. Such diodes may be either of the germanium or silicon type. The area of their *P-N* junction amounts to several square microns.

A point-contact diode is shown schematically in Fig. 8-45. A sharp-ended tungsten wire bears against the surface of

an *N*-type semiconductor which has a *P*-region near the point of contact with tungsten, thus, a *P-N* junction is formed. The lower surface of the semiconductor crystal is alloyed with a metal current collector.

The so-called *tunnel diodes* which are special semiconductor devices are of great interest for ultra-high frequency applications. We will only briefly describe these diodes. Germanium (or sometimes gallium arsenide) used for tunnel diodes is

doped with *high concentrations of impurities*. In these diodes the *P-N* junction is many times thinner than usual and, consequently, field intensity will be very high. Under these conditions the *tunnelling effect* occurs in the diode, i. e., *electrons tunnel through the potential barrier* almost instantly without any change in their energy. Electrons penetrate the *P*-region only when there are

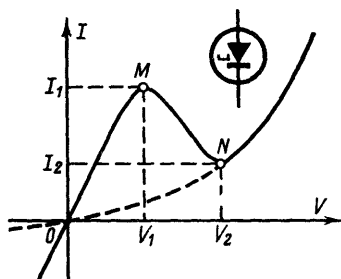


Fig. 8-46. Characteristic and circuit symbol of tunnel diode

unfilled energy levels in this region.

These phenomena give rise to a peculiar shape of a tunnel diode characteristic curve (Fig. 8-46). Owing to the tunnelling effect, current rises abruptly near the point *O* both in the forward and reverse directions at external voltages equal to small fractions of a volt. Forward tunnelling current rises up to point *M* until all free energy levels are filled with electrons. Then tunnelling current decreases inspite of the increase in voltage. Current starts to rise again at point *N* but now at the expense of electrons which are capable of overcoming the potential barrier. In other words, this portion is similar to an ordinary characteristic of a crystal diode (its low portion is shown by the dashed line passing through the origin of coordinates).

The descending portion *MN* is really remarkable—here the *resistance* of the tunnel diode is found to be *negative* since an increase in voltage corresponds to a decrease in current, i. e.,

$$R \approx \frac{I_2 - I_1}{V_2 - V_1} < 0$$

Such a semiconductor device may be used successfully for amplification and generation even at ultra-high frequencies (owing to its small sizes).

8-8. TRANSISTORS

A crystal triode is designed for amplification and generation of oscillations similarly to a vacuum triode, but their operating principles are completely different. The principle of operation of a crystal triode employing one of its basic circuit

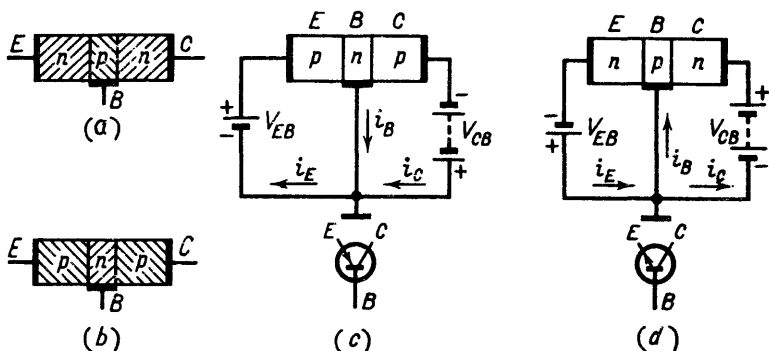


Fig. 8-47. Transistor types

(a) N-P-N transistor; (b) P-N-P transistor; (c) supply of P-N-P transistor and its circuit symbol; (d) supply of N-P-N transistor and its circuit symbol

connections may be laid down as follows: when an alternating current of a definite value feeds small input resistance of the triode, simultaneously at its output a current of approximately the same value feeds a large load resistance, thus providing power amplification. The transfer of a small input resistance to a large load resistance occurring in such a triode gave it the name "TRANSfer reSISTOR", abbreviated *transistor*.

Although the earliest use of crystal diodes in electronics goes back to the receivers of A. S. Popov and P. N. Rybkin at the end of the last century, transistors were developed only by the fifties of the present century.

The transistor basically consists of a germanium or silicon crystal and two P-N junctions (not one as in a diode) formed

by the methods described above. In essence, the transistor is made up of two diodes connected back to back. As shown in Fig. 8-47a the transistor crystal has three regions: the left-hand electron region (*N*-region) called the *emitter*, the middle hole region (*P*-region) known as the *base* and the right-hand electron region (*N*-region) called the *collector* of electrons. The transistor having such a structure is known as the *N-P-N* type transistor.

There also are transistors in which the *N*-region is between two *P*-regions as shown in Fig. 8-47b. Transistors of such a structure are called the *P-N-P* type transistors and are widely used in electronics. The majority carriers in the *P-N-P* transistor are holes.

Since the emitter of a transistor serves as a source of charge carriers (analogous to the cathode in an electron valve), the emitter region should have the highest concentration of its own majority carriers, which is achieved by appropriate doping it with impurities. The concentration of charge carriers in the collector region is less than in the emitter region since it receives, so to say, "foreign" carriers (analogous to the anode in an electron valve). The base region has the lowest concentration of its own majority carriers as it controls the motion of "foreign" carriers (analogous to the control grid in an electron valve). Each region is provided with a metal terminal (a lead) and therefore, the transistor derives its name "crystal triode" from these three terminals.

Let us now connect DC voltage sources between the emitter and base terminals, V_{EB} , and the collector and base terminals, V_{BC} , observing the following rule: a low *emitter-to-base voltage* (a fraction of a volt in low-power transistors) and a *collector-to-base voltage* (volts, sometimes over a dozen of volts) must be *forward* and *reverse (cutoff)* voltages, respectively. Fig. 8-47c illustrates the supply sources for the *P-N-P* type transistor. Note that the negative and positive terminals of the corresponding batteries are connected to the collector and emitter, respectively, so that the current flow (motion of positive charges) is directed from the emitter to the base. It is this motion that is indicated by the arrow in the circuit symbol of the *P-N-P* type transistor shown in the same figure. The supply sources for the *N-P-N* type transistor shown in Fig. 8-47d are connected in compliance with

the same requirements for the transistor junctions, i. e., the collector circuit is connected here to the battery positive terminal. Current flow (opposite in direction to electron flow) is directed from the base to the emitter and the reverse direction of the emitter arrow permits the $N-P-N$ transistor to be distinguished from the $P-N-P$ transistor in schematic diagrams.

It should be noted here that $P-N-P$ type transistors are easy to mass produce and, therefore, the majority of transistors have this structure. However, the majority carriers in the $P-N-P$ type transistors are holes which have a lower mobility than electrons. The $N-P-N$ type transistors look more promising for ultrahigh-frequency applications. Generally speaking, a transistor acts more slowly than an electron valve, as in solids electrons move slower than in vacuum. Moreover, at ultrahigh frequencies $P-N$ junctions are generally shunted by their capacitances. As a consequence, transistors enjoy less advantage as compared with purely electron devices and crystal diodes in the range of decimetre and especially centimetre waves.

Let us return to the schematic diagram shown in Fig. 8-47c and consider the directions and relationships between the currents in the connecting wires. The holes from the left P -region repelled by the positive voltage V_{EB} diffuse through the left-hand junction (low potential barrier) and move into the N -region of the base (in forward direction). The holes partially recombine here with base electrons. The base region is made, however, extremely thin (fractions of a millimetre, this is very important!) and, therefore, the great majority of holes (up to 99%) reach the right-hand junction by diffusion. The holes (being here the minority carriers) pass freely through the right-hand junction and after drifting through the right P -region are attracted by the negative voltage V_{CB} , thus producing the current i_C in the collector circuit. The charges carried by holes from the left P -region are replenished by the current i_E flowing in the emitter circuit through the voltage source V_{EB} . Finally, the base region electrons recombining with emitter holes are supplied from the external circuit, thus producing the base current i_B . The base current is approximately the difference of emitter and collector currents as follows from the process discussed abo-

ve, i.e.,

$$i_B = i_E - i_C \quad (8-19)$$

In the circuit diagram shown in Fig. 8-47d the formation of currents is similar to that described above, with the difference that electrons, not holes, move to the right from the emitter, and currents in the circuits flow in reverse directions. Therefore, the processes occurring in the *N-P-N* type transistor are self-evident.

Let us discuss in some detail the physical processes occurring in the *P-N-P* type transistor (Fig. 8-48a). If the emitter circuit is open, then the current flowing from the collector through the base will not disappear completely. While approaching the right *P-N* junction, the conduction electrons from the right *P*-region move into the *N*-region and produce the current I_{C0} shown by the dashed line in Fig. 8-48. In the collector circuit the direction of this current coincides with that of the

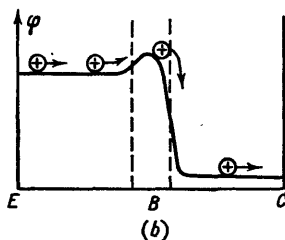
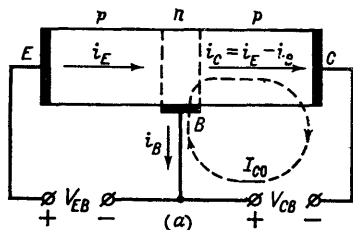


Fig. 8-48. Processes in *P-N-P* transistor

(a) currents; (b) hole flow

main current i_C , whereas in the base circuit the current i_{C0} is opposite to the main current i_B , thus the base current presents the difference of these two currents. The current I_{C0} is known as the *collector cutoff current*. It is rather low as it is produced by minority carriers of the right *P*-region. The base current i_B is generally greater than I_{C0} and its direction is determined by the holes moving from the base region. However, the current I_{C0} must be taken into account since it rises abruptly with temperature (approximately doubles for each 10°C). This should be considered when providing temperature stability of transistor equipment.

Let us trace the main path of majority carriers (holes) in the *P-N-P* type transistor as shown in Fig. 8-48b. The holes

move from the emitter through the emitter junction (whose potential barrier is reduced by the voltage V_{EB}) and then through the collector junction (whose electric field permits minority carriers to pass freely), and finally the holes are attracted to the negative collector electrode. This process is remarkable by the following: *the value of collector current is determined* by the flow of holes through the left junction, i. e., *by the voltage V_{EB}* , since only several and even a fraction of a per cent of the over-all amount of holes are branched off into the base circuit, whereas field intensity at the right junction is always sufficient for attracting holes into the collector region, and therefore voltage at the collector has practically no effect on collector current. The last property is similar to the influence of anode voltage on anode current in a pentode.

The processes occurring in the transistor and its properties are illustrated graphically by its characteristic curves. In the preceding sections we discussed the characteristic curves of anode and grid currents in a vacuum triode as functions of the voltages being applied. The characteristic curves of a transistor also represent current-voltage relationships.

Let us discuss a simple (but not always the best) transistor circuit similar to those shown in Figs. 8-47 and 8-48, which may be referred to as a *common-base configuration* since the base terminal is the common (grounded) point for all three electrodes. Such a circuit is given in Fig. 8-49a. All circuit characteristic curves may be taken by connecting the instruments as shown in the figure and measuring the currents as the voltages are changed by steps. The emitter circuit is called the *input circuit* as a signal source will be connected into it. The collector circuit is termed the *output circuit* since a load will be connected into it. Relationships $i_E = \varphi(u_{EB})$ and $i_C = f(u_{CB})$ are called, respectively, *input* and *output characteristics* of a common-base transistor.

The input characteristic (Fig. 8-49b), i. e., the relationship between emitter current and input (emitter-to-base) voltage for a constant collector-to-base voltage is similar to the characteristic curve of a forward-biased crystal diode (see Fig. 8-43b) since a forward biasing voltage is applied to the left *P-N* junction. If $u_{CB} = 0$, i. e., if the collector is shorted to the base, we obtain precisely the diode characteristic curve passing

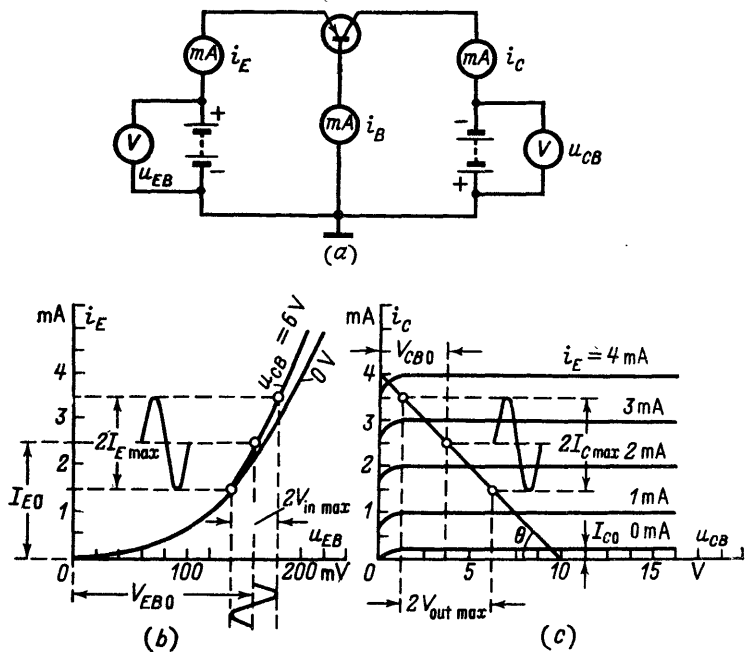


Fig. 8-49. Common-base transistor

(a) schematic diagram; (b) input characteristics; (c) output characteristics

through the origin of coordinates. When collector-to-base voltage is increased, the input characteristic is shifted slightly upwards since the collector voltage, as already stated above, has practically no effect on the emitter junction. For this reason, only one input characteristic is sometimes given in manuals for a common-base transistor.

Let us examine output characteristic curves (Fig. 8-49c) proceeding from the lowest one corresponding to $i_E = 0$, i. e., when the emitter circuit is open. This is a characteristic curve for the collector cutoff current I_{C0} ; it runs from the origin of coordinates and represents a current flowing through the diode in the reverse direction. At normal temperatures collector cutoff current in receiver transistors amounts to several microamperes. The curves located above the collector cutoff current characteristic curve represent a family

of output characteristics for various values of constant emitter current. As is known, *collector current is proportional and is almost equal to emitter current* (except for a small fraction of charge carriers "intercepted" by the base). Therefore, the characteristic curves are linear, nearly horizontal and equally spaced one from another. The slight rise of the nearly horizontal portions of the characteristic curves is caused by a spread of the right-hand junction with growth of the cutoff voltage V_{CB} and, consequently, the lessening base thickness leading to a smaller percentage of charge carriers which manage to recombine in the base region. Note that plotting the voltage V_{CB} to the right we do not indicate its negative sign, this, however, should be remembered when dealing with $P-N-P$ transistors.

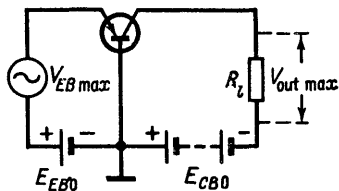


Fig. 8-50. Common-base transistor amplifier with a load resistor

What are the advantages and disadvantages of the common-base circuit? Let us connect a source of signal voltage $V_{EB \max}$ (Fig. 8-50) with the amplitude $V_{in \max}$ into the input circuit after selecting the initial positive bias V_{EB0} (Fig. 8-49b) so as to operate on the linear portion of the input characteristic curve (thus, eliminating nonlinear distortions). Under these conditions an AC component with amplitude $I_E \max$ appears in the emitter circuit. With input voltage equal to several millivolts this current will prove to be equal to several milliamperes. This indicates the high input conductivity of the common-base circuit. In Fig. 8-49b we have approximately $G_{in} = I_{in \max} / V_{in \max} = 1 \times 10^{-3} / 20 \times 10^{-5} = 0.05$ S. In practice, G_{in} is not less than hundredths of a millisiemens. In other words, the *input resistance* R_{in} amounts to only *dozens of ohms*.

Let us connect a load resistor R_L into the collector circuit of the diagram shown in Fig. 8-50 and assume the signal voltage across R_L and the power in it to be the *useful output of a common-base transistor amplifier*. It will be recalled that flat output characteristic curves (analogous to anode characteristic curves of a pentode) indicate a *high internal (output) resistance* R_{out} of the semiconductor device. This is quite natu-

ral since the collector junction with reverse bias has a really high resistance (dozens and hundreds of thousands of ohms). Therefore, the resistor R_l may be chosen to be more than 1 k Ω provided the DC voltage drop across the collector is permissible. In Fig. 8-49c the line of load is shown at the angle θ corresponding to $R_l=2500\ \Omega$ with the battery voltage $E_{CB}=10\text{ V}$ (see Fig. 8-25). The amplitude of alternating collector current $I_{C\text{ max}}$ will be apparently almost equal to that of the emitter current $I_{E\text{ max}}$. This may be expressed more precisely in the formula

$$\begin{aligned} I_{C\text{ max}} &= \alpha I_{E\text{ max}} \\ I_{B\text{ max}} &= (1 - \alpha) I_{E\text{ max}} \end{aligned} \quad (8-20)$$

where $\alpha < 1$ is that portion of charge carriers which moves from the emitter through the base and reaches the collector. The parameter α is defined as the emitter-to-collector current gain of a common-base transistor or simply the *common-base current gain*. The value of α may vary from 0.95 (for the worst transistors) to 0.998 (for the best transistors). For practical purposes, α is generally rounded off to unity.

Voltage gain in the common-base transistor amplifier will be evidently more than unity, if the output current nearly equals the input current and the load resistance exceeds the input resistance, i. e.,

$$K = \frac{V_{out\text{ max}}}{V_{EB\text{ max}}} = \frac{I_{C\text{ max}} R_l}{I_{E\text{ max}} R_{in}} = \alpha \frac{R_l}{R_{in}} \quad (8-21)$$

Thus, we have explained physically and determined quantitatively signal voltage amplification in the common-base transistor amplifier. If we assume, for instance, that $\alpha=0.98$, $R_{in}=50\ \Omega$ and $R_l=2500\ \Omega$, then the voltage gain will be

$$K = 0.98 \frac{2500}{50} \approx 50$$

Power gain K_P will prove to be nearly the same as the voltage gain since

$$K_P = \frac{I_{C\text{ max}}^2 R_l}{I_{E\text{ max}}^2 R_{in}} = \frac{V_{out\text{ max}}^2 G_l}{V_{EB\text{ max}}^2 G_{in}} = \alpha^2 \frac{R_l}{R_{in}}$$

Now the common-base transistor amplifier may be characterized rather fully. It has a very low input resistance and a

high output (internal) resistance. This means that it is suitable for applications where the signal source has a low internal resistance (in Fig. 8-50 it was assumed to be zero) and the load resistance can be chosen sufficiently high (thousands or at least many hundreds of ohms).

The current gain was proved to be less than unity ($K_{cur} = \alpha$) and the power gain was equal roughly to the voltage gain. The output characteristic curves of the common-base transistor amplifier are flat and parallel to each other, thus providing operation without linear distortions (this will be explained in more detail in the next chapter).

The transistor may also employ another circuit connection which permits signal voltage and current to be amplified. This circuit configuration is termed a *common-emitter circuit* (Fig. 8-51a). The common-emitter connection of a transistor is similar to the connection of a vacuum triode discussed somewhere above. In fact, the emitter, base and collector are connected in much the same way as the triode cathode, grid and anode, respectively. The signal source having the voltage

$V_{BE \max}$ is connected into the base circuit in addition to the initial bias source E_{BE0} . The load (for instance, the resistor R_L) is connected into the collector circuit in series with the supply source E_{CE} .

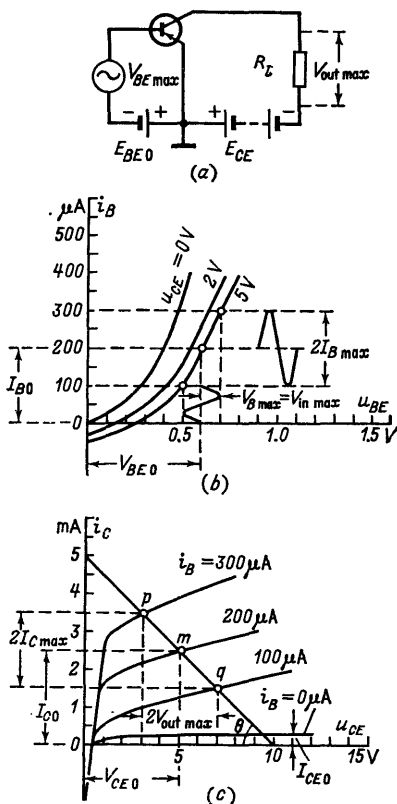


Fig. 8-51. Common-emitter transistor amplifier with a load resistor; (b) input characteristic curves; (c) output characteristic curves

Let us consider the characteristic curves of the common-emitter transistor under static conditions, i. e., when the terminals of the signal source and the resistor R_i are shorted.

The input characteristic relates base current to base-to-emitter voltage for a constant collector-to-emitter (output) voltage (Fig. 8-51b). In low-power transistors base current amounts to a *fraction of a milliampere* (i. e., microamperes). The characteristic curve corresponding to $u_{CE}=0$ (collector and emitter are shorted) is similar to the characteristic curve of a forward-biased crystal diode. If now voltage u_{CE} is turned on, the base current of the new characteristic curve will decrease at the same voltages across the base since the collector will intercept charge carriers moving from the base region. With a small voltage applied to the base, base current even reverses its direction. This is explained by the fact that the collector cutoff current flowing through the base wire in the opposite direction to the base current (see Fig. 8-48a) becomes dominant in the base circuit. With a further increase in the voltage u_{CE} (5 V instead of 2 V in Fig. 8-51b), the input characteristic curve is *shifted only slightly*. Therefore, we may limit ourselves to a single input characteristic curve taken at a normal collector voltage. The right-hand curve taken at $u_{CE}=5$ V may serve as such a characteristic curve in the example shown in Fig. 8-51b.

The output characteristic of the common-emitter transistor relates collector current to collector-to-emitter voltage for a constant base current. Instruments used for taking output characteristic curves are the same as shown in Fig. 8-49a. The lower curve in the family of output characteristics shown in Fig. 8-51c corresponds to $i_B=0$ which means that the base circuit is open. Under these conditions *open-base emitter current* I_{CE0} (of about hundreds of microamperes for receiver transistors) flows from the emitter to the collector. This current is greater than conventional diode reverse current since collector-to-emitter voltage drops partially across the emitter junction in the forward direction.

When $i_B > 0$, the output characteristic curve runs above the one for $i_B=0$ (Fig. 8-51c). In fact, when base current is increased, the emitter current (being many times the base current) also increases. Simultaneously, the collector current increases with emitter current. On the operating portion of

the output characteristic curve collector current will somewhat increase as the voltage u_{CE} is increased and the current i_B is hold constant. In fact, if the voltage across the collector junction is increased, the base thickness is reduced, thus the base current will go down. But the output characteristics are plotted for $i_B = \text{constant}$, and one has to raise the bias voltage E_{BE} so as to maintain the base and emitter currents constant. When the base current is given a next step rise, a new characteristic curve running above and parallel to the previous characteristic is obtained. By this means a family of the output characteristic curves shown in Fig. 8-51c can be obtained. Their operating portions are steeper than those of output characteristics for the common-base transistor (see Fig. 8-49c). This is an indication that the output (internal) resistance of the common-emitter transistor is less than that of the common-base transistor. It amounts generally to *thousands or several dozen of thousands of ohms*. However, the input resistance of the common-emitter transistor is hundreds and even higher than a thousand ohms as can be seen from Fig. 8-51b. It is by these parameters that the common-emitter circuit is distinguished basically from the common-base circuit.

Before considering the amplifying properties of the common-emitter transistor, attention should be given to the initial portions of its output characteristic curves. As collector-to-emitter voltage u_{CE} drops down to zero (with the base circuit being closed), collector current sharply falls to zero and then starts to flow in the reverse direction. The fact is that the forward bias voltage supplied by the base circuit source and applied to the collector junction becomes dominant under these conditions. Collector reverse current is practically zero when a transistor operates as an amplifier and, therefore, the fact of current cutoff (and the value of cutoff voltage at the collector) becomes of great importance.

Let us now return to the amplifying properties of the common-emitter circuit (Fig. 8-51a) capable of *amplifying the signal current*. In fact, if the load resistance were zero (the output terminals short-circuited), the collector alternating current $I_{C \text{ max}}$ would amount to the fraction α of the emitter alternating current $I_{E \text{ max}}$. At the same time the input (signal) current is the base current which, as we know, amounts to

the fraction $(1-\alpha)$ of the emitter current. Consequently,

$$I_{B \max} = (1-\alpha) I_{E \max} = \frac{1-\alpha}{\alpha} I_{C \max}$$

Hence, with the output terminals short-circuited the current gain will be

$$\frac{I_{C \max}}{I_{B \max}} = \frac{\alpha}{1-\alpha} = \beta \quad (8-22)$$

The parameter β is defined as the *current gain of the common-emitter transistor*. In fact, current is amplified in this circuit since $\beta > 1$. If, for instance, $\alpha = 0.98$, then $\beta = 0.98/(1-0.98) = 49$. Current amplification can be explained quite easily, namely, in the common-emitter circuit the signal controls great collector current when base current is small! In actual amplifiers (having a load resistor), however, current gain will be less than β .

Moreover, the common-emitter circuit permits signal voltage to be amplified as well. Let us connect the load resistor R_l of resistance equal to (for instance) 2000Ω into the collector circuit, and construct the line of load at the corresponding angle θ (see Fig. 8-51c). When signal voltage $V_{in \max} = 0.1 \text{ V}$ is applied to the input (with the initial bias corresponding to a current of $200 \mu\text{A}$ at a collector-to-emitter voltage of 5 V), an alternating current $I_{B \max}$ of about $100 \mu\text{A}$ appears in the base circuit. Under these conditions the operating point will move up and down along the portion *pmq* of the load line (i. e., between the static characteristics corresponding approximately to base currents of 300 and $100 \mu\text{A}$). The amplitude of output alternating voltage, $V_{out \max}$, will run as high as 2 V . Hence, voltage gain

$$K = \frac{V_{out \max}}{V_{in \max}} \approx \frac{2}{0.1} = 20$$

Current gain will differ from its static value, β . For the case considered (Fig. 8-51) it will approximately amount to:

$$K_{cur} = \frac{I_{C \max}}{I_{B \max}} = \frac{10^{-3}}{10^{-4}} \approx 10$$

Power gain will equal the product of voltage and current gains, i. e.,

$$K_P = K \cdot K_{cur} = 20 \times 10 = 200$$

The advantages of the common-emitter transistor amplifier become clear even from this example, though very approximate and random. In fact, this configuration represents the *basic circuit* most widely used in transistor amplifying equipment.

8-9. MINIATURIZATION OF ELECTRONIC DEVICES

Transistors are being continuously improved as regards the range of the frequencies being amplified, temperature stability of transistor parameters, power ratings, etc. New designs and even operating principles of transistors, the enumeration of which is not possible here, are now making their appearance in commercial equipment. One of the most remarkable trends, however, in the development of transistors is their *miniaturization* making it possible to reduce significantly electronic equipment sizes.

The most advanced developments in modern technology such as spacecraft or self-organizing automatic control systems require the application of sophisticated but highly reliable and small-size electronic equipment. Small weight and volume requirements imposed upon radio equipment, the more complicated missions being entrusted to radio systems, and an ever increasing number of their components—all necessitate their size reduction. This is inconsistent, however, with the requirements of high reliability since the greater the number of components in a system, the more probable is failure of at least one of them, thus resulting in failure of the entire system. Moreover, the smaller the component, the lower is its mechanical and electrical strength, i. e., the lower is the reliability of the component. Sophisticated wiring involving a great number of soldering or welding operations is expensive and also unreliable.

The need for reducing the sizes of semiconductor devices stimulated the development of microminiature devices as a whole. Recent achievements in physics and solid-state technology made it possible to implement a great number of components in a small volume, thus allowing the size of electronic devices to be reduced thousands of times as compared to conventional electron valve or even semiconductor equipment performing similar functions. Simultaneously,

automated production reduces the cost of the products and assures high reliability.

The most advanced trends in miniaturization that come into being in recent years are *thin-film* and *solid-state microelectronics*.

An integrated circuit is a microelectronic circuit which is made completely on the surface or in the bulk of a single "substrate". It consists of areas or layers functioning as conventional radio components (passive or active). *Thin-film circuits* constitute one of the categories of integrated circuits. Such circuits are composed of a common dielectric substrate which carries a combination of layers (deposited by different techniques) of conducting, semiconducting and insulating materials, and this combination forms a functional unit or even a device as a whole. However, this is precisely the problem (the most complex one in this technology) of implementing thin-film transistors which, as we remember, require the single-crystal structure of a semiconductor. Therefore, it is common practice to use discrete (outer) microtransistors in a combination with thin-film circuits. Such a combination is referred to as the *hybrid circuit*.

Another category of integrated circuits is called the *solid circuits*. In such circuits components are implemented not on the surface of an insulating substrate but in the bulk of the solid semiconductor material. Pure silicon is generally the material used for this purpose. The size of wafers cut from silicon may be comparable with that of a match head. The circuit produced in silicon is sometimes referred to as a "solid-state circuit", thus reminding that this technology is based primarily on solid-state physics.

The solid circuit is formed in a pure crystal in several processing stages which present "an intrusion" changing its structure and physical properties. This can be achieved, for instance, by diffusion of the impurities which produce electron-hole junctions capable of functioning as diodes, resistors, transistors, etc. As a result of such systematic automated processing, an ultraminiature circuit comprising active and passive elements and capable of performing independent electric functions is formed in the bulk of the semiconductor wafer. A magnified photo of a solid circuit is illustrated in Fig. 8-52. The circuit case is of 6.3 mm long and 3.2 mm

wide and houses a device consisting of many dozens of active and passive elements.

Thus, microminiaturization opens up ways for solving inconsistency between the demand for complex electronic equipment involving a great number of components, on one hand, and the necessity for reliable operation of this equipment which should have small size and weight and be mass produced by automated techniques, on the other hand. We cannot state, of course, that all radio equipment will be miniaturized in future. Naturally, high-power transmitters and TV picture or camera tubes cannot be reduced much in size. But low-power radio electronics and first of all electronic computer developments depend heavily upon the achievements in microcircuit and integrated circuit technology. Solid circuits, among other things, represent a very high but not an ultimate level of microminiaturization. It is significant that the progress in the efficiency of scientific and research work is closely associated with the achievements in microelectronics (logic devices, computers, etc.).

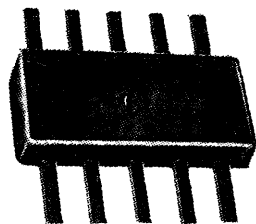


Fig. 8-52. Solid circuit

8-10. CATHODE-RAY TUBES

Cathode-ray tubes (CRT) are electron devices designed to convert electric signals into visual (light) images. Such devices find extensive application in TV receivers, radar displays, instruments and many other fields of radio electronics. In all CRT an electron beam (ray) having a small cross section and relatively great length (in some devices dozens of centimetres) is formed and deflected by an electric or magnetic field.

The formation of a thin beam (focusing an electron flow) and its deflection may be accomplished owing to interaction between moving electrons and an electric or magnetic field. Accordingly, cathode-ray tubes used in electronics are of the following types: electrostatic cathode-ray tubes (where beam focusing and deflection are performed by an electric field);

magnetic-deflection cathode-ray tubes (where the electron beam is focused and deflected by a magnetic field); combined-deflection tubes (where beam-focusing is electrostatic and deflection is magnetic or vice versa).

We first look at an electrostatic cathode-ray tube designed for scientific experiments and electronic equipment used for visual representation of various electrical processes. For instance, such a tube is the main part of a cathode-ray oscilloscope.

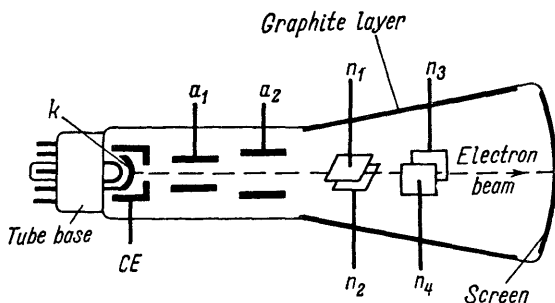


Fig. 8-53. Construction of electron-ray tube with electrostatic control

The simplest oscilloscope CRT (Fig. 8-53) is enclosed in a glass envelope. The inside of the envelope face is coated with a layer of a fluorescent material (phosphor) capable of emitting light when struck by electrons. This layer serves as a screen on which an electrical process is made visible for viewing.

In its neck (at the opposite end of the tube) the envelope has an *electron gun*—a combination of indirectly heated cathode *k* and three cylindrical electrodes centred along the tube's axis—which forms the electron beam. The leftmost is the control electrode (*CE*), also known as modulating grid, which surrounds the cathode and permits the electrons beam to pass only through an opening in its end face. The density of electron beam and, consequently, the brightness of the fluorescent screen can be varied by changing the negative potential applied to the control electrode with respect to the cathode.

The control electrode is followed by two cylindrical anodes provided with diaphragms (i. e., partitions having central apertures). A DC voltage not exceeding 300—400 V is applied to anode a_1 with respect to the cathode, and anode a_2 is supplied with a DC voltage ranging from 600 to several thousands of volts. Electrons are accelerated by the electric field of the anodes and pass through them to the fluorescent screen. Owing to the presence of diaphragms and the action of the longitudinal electric field formed between the anodes, the electrons are focused into a narrow beam which produces a bright spot at the centre of the screen.

On its way from the second anode to the screen the electron beam passes in succession two pairs of deflecting plates mounted at right angle to each other and having leads to external pins in the tube base. If no voltage is applied across the plates of one pair, that pair has no influence on the electron beam. With a voltage applied to the deflecting plates, a transverse electric field is produced between them causing beam deflection, thus displacing the electron spot over the screen. Therefore, these plates are called deflecting. The cathode-ray tube is mounted so that the pair of plates n_1 and n_2 can deflect the spot vertically and the pair of plates n_3 and n_4 horizontally.

All the electrodes shown in Fig. 8-53 are connected to pins in the tube base. The pins are inserted into a special panel, when the CRT is connected in a circuit. Secondary emission from the screen is used in the CRT to prevent the screen from being charged negatively by electrons, otherwise the negative charge on the screen would repel arriving electrons. The secondary electrons are attracted by a grounded layer of graphite coated on the inside of the tapered portion of the envelope; this layer is at a positive potential with respect to the screen.

The schematic diagram given in Fig. 8-54 illustrates one of the many possible ways of applying external voltages to the oscilloscope tube. The positive pole of the high-voltage power supply E is connected (through a resistor) to anode a_2 and grounded through the oscilloscope chassis. One deflecting plate of each pair is connected to anode a_2 . Anode a_1 is connected to the slider of potentiometer R_1 which serves to adjust the anode positive potential with respect to the cathode, thus providing focusing of the electron beam, i. e., the mi-

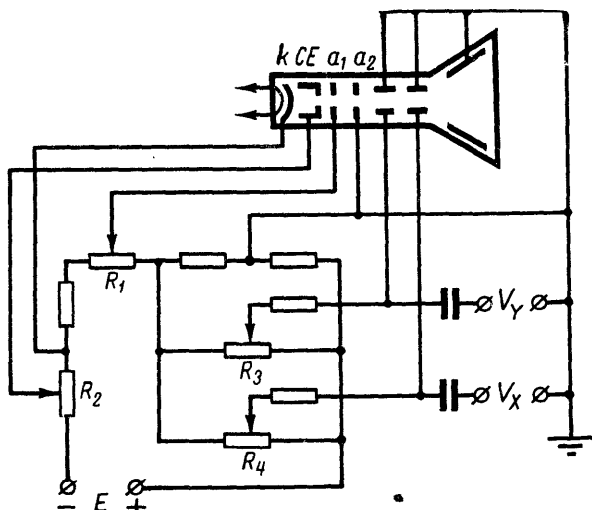


Fig. 8-54. Connection of oscilloscope cathode-ray tube

nimum spot size on the screen. The control electrode is at a negative potential with respect to the cathode. Spot brightness is adjusted by changing the position of the slider of potentiometer *R*₂. Vertical deflection (*V*_Y) and horizontal deflection (*V*_X) voltages are applied to the deflecting plates which are connected to potentiometers *R*₃ and *R*₄ used for displacing the spot over the screen.

Deflection sensitivity of the cathode-ray tube is the displacement (in millimetres) of the spot at the screen per volt applied between the deflecting plates. Conventional oscilloscopes have deflection sensitivity of 0.1 to 0.5 mm/V. Consequently, noticeable deflection is obtained with voltages in the order of dozens of volts. If it is necessary to detect the behaviour of lower voltages, they are first amplified and then applied to the deflecting plates.

The so-called linear scanning (sweep) allowing periodic electric processes to be visualized on the tube screen is the most important among many of the practical applications of an oscilloscope. For this purpose, a scanning voltage *V*_{scan} in the form of periodic saw-tooth pulses supplied from a special auxiliary generator is applied to the horizontal

deflecting plates (X axis). In Fig. 8-55 this voltage is shown with the time scale directed downward. This voltage rises linearly with time (section OA) causing the spot to move horizontally over the screen. Then it drops abruptly (section AB), and the spot suddenly snaps back to its initial position. If the repetition frequency of the saw-tooth pulses is more than 20 times per second, an observer sees a bright horizontal trace whose length is proportional to the saw-tooth voltage amplitude, AB , and whose width is equal to the spot diameter.

Simultaneously, a periodic voltage V_p corresponding to the electric process of interest is applied to the vertical deflecting plates (Y axis). In Fig. 8-55 this process is represented by the periodic pulses shown upper right. If the repetition frequency of saw-tooth pulses is equal to that of the process, the spot will experience one vertical deflection during one left-to-right (horizontal) sweep (scan). These vertical deflections will be repeated on the screen, thus superimposing and producing a visual representation of the pulses as shown in Fig. 8-55.

If voltage pulses with a known repetition period are applied to the control electrode (in the presence of the voltages mentioned above), bright time marks will appear on the oscillogram of the process being examined. These time marks may be used for determining the length of the pulses being studied and their repetition period. As far as the scanning (sweep) generator is concerned, its function is to produce saw-tooth voltage with variable frequency and to maintain synchronism with the process being studied.

To understand the operating principle of magnetic-deflection cathode-ray tubes, let us first familiarize ourselves with the behaviour of an electron in a magnetic field. A magnetic field exerts no effect on a stationary electron. However, if

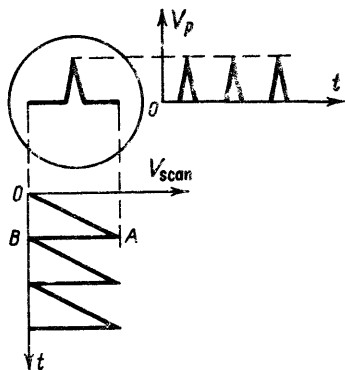


Fig. 8-55. Scanning of periodic process on oscilloscope screen

the electron moves in the magnetic field, a force arises which acts on it similar in its nature to the force which acts on a current-carrying conductor in a magnetic field. Therefore, when the electron moves along the magnetic line of force, the force is equal to zero (as in the case of a current-carrying conductor).

Consider the case when an electron moves in uniform magnetic field in the plane perpendicular to the direction of magnetic lines of force. Then the electron is acted upon by force F whose direction is determined by the left-hand rule. As will be recalled, the direction of current is opposite to that of electron motion (Fig. 8-56). The force is directed perpendicularly to the electron velocity vector v and the magnetic field intensity vector H . This force is determined by

$$F = eHv$$

where e is the electron charge.

Since the force is perpendicular to the direction of electron motion, it performs no work, i. e., it does not increase or decrease electron energy. This force, however, changes the trajectory of the electron making it to deflect from the initial direction.

The electron path, initially a linear one, is turned into a circle whose radius can easily be found by considering that F is a centripetal force equal to mv^2/ρ where m is the mass of the electron, v , its velocity and ρ , the radius of the circle described by the electron:

$$eHv = \frac{mv^2}{\rho}, \text{ i. e., } \rho = \frac{mv}{eH} \quad (8-23)$$

As m/e is constant, radius ρ will be the smaller, the higher the field intensity and the less the electron velocity. Slow electrons are deflected (twisted) more in a high-intensity magnetic field than fast electrons in a weak magnetic field.

The magnetic field influence on the electron flow is used extensively in engineering. In particular, a magnetic field may be used in cathode-ray tubes both for focusing and deflecting the electron beam.

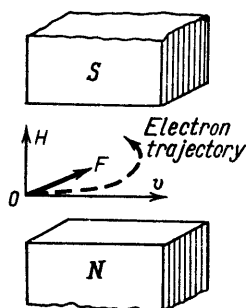


Fig. 8-56. Motion of an electron in transverse magnetic field

Magnetic focusing is accomplished with the aid of a focusing coil (*FC*) mounted coaxially around the neck of the cathode-ray tube (Fig. 8-57*a*). The tube has a control electrode and an anode. In some instances, the graphite layer coated on the inside of the envelope and to which a positive potential is applied is used as the anode. Electrons moving along the tube axis are unaffected by the coil field and pass straight to the screen. Electrons deflecting from the tube axis acquire

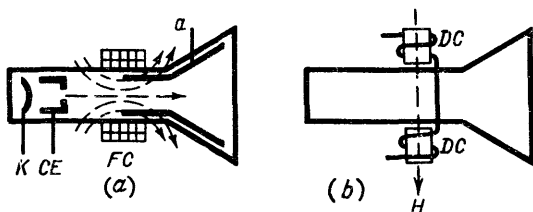


Fig. 8-57. Electron-ray tube with magnetic control
(a) focusing principle; (b) deflection principle

a velocity component perpendicular to the magnetic lines of force. These electrons are "twisted" towards the tube axis. Magnetic field intensity should be selected so that electrons can converge at one point on the screen surface. This is achieved by adjusting the direct current supplying the focusing coil.

Magnetic deflection of an electron beam is carried out with the aid of two accordant- and series-connected deflecting coils (*DC*) which produce a magnetic field perpendicular to the electron flow (Fig. 8-57*b*). An electron passing through the magnetic field of these coils has its trajectory curved but travels linearly along the changed direction on leaving the magnetic field. The higher the current in the coils the greater the deflection of the spot over the screen. To deflect the electron beam horizontally and vertically, the cathode-ray tube is provided with two pair of coils located at right angles to one another.

Magnetic focusing and magnetic deflection are used for most modern kinescopes, i. e., TV picture tubes.

9-1. PURPOSE AND CLASSIFICATION

An amplifier is a device designed to produce electrical oscillations of higher power than that of the signals controlling its operation. The energy of electrical oscillations produced by the amplifier is drawn from a supply source. An electron valve or transistor controlled by the signal being amplified serves as a converter of the energy supplied by a DC power source into that of amplified signal.

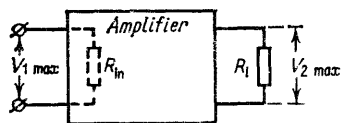


Fig. 9-1. Block diagram of amplifier

The amplifier terminals to which a control signal is applied are referred to as the input terminals or simply the *input*.

The source of the signal to be amplified (a signal source) delivers an alternating voltage $V_{1 \max}$ (Fig. 9-1) to the amplifier input. The amplifier itself presents some resistance R_{in} , called the *amplifier input resistance*, to the signal source.

Of course, the input resistance serves only as a measure for the load presented by the amplifier to the signal source when the latter is connected to the grid-cathode terminals of an electron valve or to the *P-N* type junction of a semiconductor device and delivers the power P_{in} to the amplifier input. Using the resistance R_{in} instead of an actual load, we can determine the value of this resistance from the equation

$$P_{in} = \frac{V_{1 \max}^2}{2R_{in}} \quad (9-1)$$

Assume that peak input voltage $V_{1\text{ max}} = 1\text{ mV} = 1 \times 10^{-3}\text{ V}$ and the input power $P_{in} = 10^{-10}\text{ W}$, then the input resistance

$$R_{in} = \frac{V_{1\text{ max}}^2}{2P_{in}} = \frac{10^{-6}}{2 \times 10^{-10}} = 5000\text{ ohms}$$

The input resistance is usually chosen as high as possible in order to reduce signal power consumption, however, sometimes, it is impossible to build an amplifier with a high input resistance.

Resistor R_l called a load resistor is connected to the amplifier output terminals. Alternating voltage $V_{2\text{ max}}$ is developed across this resistor, so that the power delivered by the amplifier to it is

$$P_{out} = \frac{V_{2\text{ max}}^2}{2R_l} \quad (9-2)$$

Performance of the amplifier is estimated by the *power gain* defined as

$$K_P = \frac{P_{out}}{P_{in}} \quad (9-3)$$

As we shall see later, this is only one of many performance characteristics of an amplifier.

One valve or one transistor cannot usually provide the gain required by some device. Therefore, the amplifier is generally made up of several amplification stages. An *amplification stage* (or an *amplifier stage*) is a combination of an electron valve (or transistor) with all associated circuit components which determine its performance. The signal amplified by the first stage is applied to the input of the second amplifier stage, etc. In the majority of cases supply sources are common for all amplifier stages.

In broadcast receivers the function of the last amplifier stage termed the *output stage* is to deliver the output power required for normal operation of the load, for instance, a loudspeaker, i. e., to ensure the required voltage and current in the load. All the preceding stages are designed to provide the resultant signal sufficient for the last amplifier stage to deliver the required power to the load. These stages may be called *preamplifier stages*. If we deal with amplifiers which employ electron valves operating without any grid currents

(in the region of grid negative bias), then such valves, as we know, consume negligible power at their input (see Fig. 8-22 or 8-24). In such cases, performance of an amplifier stage should be estimated by the *voltage gain* defined as the ratio of the stage output voltage to the input voltage

$$K = \frac{V_{2 \max}}{V_{1 \max}} \quad (9-4)$$

Certainly, output power of each stage should be always greater than input power. It is precisely this fact that distinguishes an *active* stage (amplifier) from a *passive* network (such as a conventional transformer) which may change signal voltage or current without adding power.

The power gain is related to the voltage gain. In fact, if the values of P_{in} and P_{out} from formulas (9-1) and (9-2) are substituted into equation (9-3), we obtain

$$K_p = \left(\frac{V_{2 \max}}{V_{1 \max}} \right)^2 \frac{R_{in}}{R_l} = K^2 \frac{R_{in}}{R_l} \quad (9-5)$$

When load resistance is small as compared with input resistance ($R_l < R_{in}$), power gain may be obtained even in those amplifier stages for which the voltage gain is less than unity. Among these stages are the output stages which operate low-resistance loudspeakers. In this case, the required power is provided due to the high current flowing in the load.

As distinct from valve amplifiers, transistorized amplifiers cannot operate without current in their input circuit. Therefore, an important operating characteristic for such amplifiers is also the *current gain* defined as

$$K_c = \frac{I_{2 \max}}{I_{1 \max}} \quad (9-6)$$

where $I_{2 \max}$ = the amplitude of current in the load resistance, and

$I_{1 \max}$ = the amplitude of current in the input circuit

It is easy to comprehend that the power gain of an amplifier (with effective output and input resistances) is equal to the product of the voltage and current gains, provided the input and output impedances are pure resistances. Thus, we have

$$K_p = K K_c \quad (9-7)$$

For a multistage amplifier the total gain is equal naturally to the product of the like gains, namely,

$$\begin{aligned}K_T &= K_1 K_2 K_3 \dots \\K_{cT} &= K_{c1} K_{c2} K_{c3} \dots \\K_{PT} &= K_{P1} K_{P2} K_{P3} \dots\end{aligned}$$

When computing the total gain of such an amplifier, it is convenient to replace the multiplication of the corresponding numerical values by the addition of their logarithms. A *decibel* (dB) is usually taken as a logarithmic unit of power. One decibel is one-tenth of the common logarithm. The power gain expressed in decibels is 10 times the common logarithm of the ratio of the power values, i. e.,

$$K_{[\text{dB}]} = 10 \log_{10} \frac{P_{out}}{P_{in}}$$

For instance, power amplification by 10 and 1000 times corresponds to 10 and 30 dB.

Thus, we can write that

$$K_T [\text{dB}] = K_1 [\text{dB}] + K_2 [\text{dB}] + \dots$$

Taking the logarithm of equation (9-5), we obtain

$$K_{[\text{dB}]} = 10 \log_{10} \left[\left(\frac{V_2 \max}{V_1 \max} \right)^2 \frac{R_{in}}{R_l} \right] = 20 \log_{10} \frac{V_2 \max}{V_1 \max} + 10 \log_{10} \frac{R_{in}}{R_l}$$

For the particular case when $R_{in} = R_l$ we have

$$K_{[\text{dB}]} = 20 \log_{10} \frac{V_2 \max}{V_1 \max}$$

This means that if we deal with voltage gain, we should express it in logarithmic units as follows:

$$K_{[\text{dB}]} = 20 \log_{10} K$$

For instance, when voltage has increased by 6 dB, we say that the voltage has doubled

$$K = 10^{K_{[\text{dB}]/20} = 10^{6/20} \approx 2$$

A decrease of voltage by 3 dB corresponds to its drop by about $\sqrt{2} = 1.41$ times or down to 0.7 of its initial magnitude.

This level (0.7) determines the bandwidth of an oscillatory circuit (see formula 2-51).

Thus, the *gain* is the first performance characteristic of an amplifier stage or the amplifier as a whole. The second performance characteristic is referred to as the *passband*. (However, an equivalent term "bandwidth" is very frequently used.) This is a range of frequencies within which the amplifier gain is approximately constant and only at its boundaries (edges) may it decrease a specified number of times.

From the viewpoint of passband arrangement on the frequency scale, amplifiers may be classified (rather roughly) into two large classes: *primary-signal amplifiers and radio-signal amplifiers* (RF amplifiers). Primary signals (or control signals) are electrical processes which represent the intelligence being transmitted. Among these are telegraph signals, i. e., DC current pulses corresponding to the alphabetic code; telephone signals, i. e., audio frequency currents representing speech or music being converted by a microphone, and finally television signals, i. e., current pulses representing a scene being transmitted with the aid of a television camera tube and scanning devices associated with it. In radio transmitters primary signals are amplified and then control (perform *modulation*) high-frequency oscillations generated in the transmitter. In radio receivers primary signals are restored from the transmitted oscillations by means of the *detection process*. On being duly amplified, these primary signals control a telegraph receiver, loudspeaker or kinescope.

A transmitting antenna radiates radio signals into space, whereas a receiving antenna intercepts radio signals, i. e., radio-frequency oscillations which have been controlled by primary signals. It will be recalled that only electromagnetic fields of high frequencies (HF) can be radiated into space. Radio signals must be amplified—in a transmitter (from the modulating stage to the antenna) to increase radiation power, and in a receiver (from the antenna to the detector) to boost radio-signal power up to the level required for effective detection. This indicates the need for radio signal amplifiers.

What is the principal difference between primary-signal and radio-signal amplifiers? The difference consists in the fact that the *frequency band* (or *spectrum*) of *primary signals* begins almost from zero frequency but most often from dozens

or hundreds of hertz. This band extends up to several kilohertz for telephone (audio) signals and up to several megahertz for television signals. For instance, the frequency band of primary signals used in commercial telephony extends from 300 to 3400 Hz, whereas in radio communication the primary signals representing music require a bandwidth at least from 100 to 10,000 Hz. The bandwidth of television primary signals (video signals) extends from dozens of hertz to 5 or 6 MHz. In some instances, primary signal amplifiers are called low- or audio-frequency amplifiers but, as we see, such a name cannot be used for television where primary signal frequencies may be as high as several megahertz which are essentially high frequencies. Therefore, the term "low-frequency amplifiers" we shall use only when dealing with *audio-frequency equipment*.

All primary-signal amplifiers feature a very high ratio of the upper to lower boundary frequencies. This ratio will be approximately a hundred in radio broadcasting and dozens of thousands in television. Resonance phenomena emphasizing a relatively narrow band of frequencies are completely undesirable in such amplifiers. Therefore, from the viewpoint of their circuit, the primary-signal amplifiers should be of the *aperiodic type*. Audio-frequency amplifiers and video amplifiers may be considered as *narrow-band* and *wide-band amplifiers*, respectively.

RF amplifiers (sometimes called *HF amplifiers*) have their passband arranged around some centre frequency which corresponds, for instance, to the transmitter wavelength (carrier frequency). In its absolute value a radio-signal bandwidth (i. e., spectrum width) in most cases exceeds by approximately twice the primary signal bandwidth, and the ratio of the upper to lower frequencies slightly exceeds unity. Thus, for radio broadcasting at a carrier frequency of 10^6 Hz ($\lambda = 300$ m) the radio-signal spectrum extends from $(10^6 - 10^4)$ to $(10^6 + 10^4)$ Hz, hence the ratio of spectrum limit frequencies (upper and lower boundary frequencies) is equal to $1,010,000 / 990,000 \approx 1.02$ and their difference is only 2 per cent. This fact explains the basic difference between radio-signal and primary-signal amplifiers. Since their *relative bandwidth* is so narrow, radio-signal amplifiers are, as a rule, of the *tuned type*, i. e., they contain resonant circuits. At the same

time, from the point of view of their *absolute bandwidth*, tuned amplifiers may be also subdivided into narrow-band amplifiers (such as used in radiotelephony) and wide-band amplifiers (those used in TV equipment).

We shall now discuss primary-signal amplifiers, first for signals of audio-frequency (AF), i. e., narrow-band amplifiers, and then for wide-band signals. RF amplifiers are considered in the chapter "Receivers", i. e., after the description of the radio-signal formation in a transmitter, including the spectra of such signals.

9-2. AUDIO-FREQUENCY AMPLIFIERS. GENERAL

An AF amplifier is given a name depending on the type of the active device used (triode, pentode or transistor) and the kind of the load connected into its output circuit (resistor, choke, transformer, etc.). For instance, there exist valve or transistor amplifiers and resistance- or transformer-coupled amplifiers, etc. Output (final) amplifier stages loaded generally with a loudspeaker occupy a particular position in this classification. As far as the amplifiers used in TV equipment are concerned, they are in most cases of the resistance type with their own distinctive features.

Let us now consider the AF preamplifier stages.

The requirements imposed upon an AF amplifier may be of a design or economical nature. Here, we shall dwell upon requirements residing in the fact that an amplifier should provide sufficient gain within the desired band of audio frequencies with a specified distortion level.

The wider the frequency band, the more difficult to provide uniform amplification of all frequencies within this band. Non-uniform amplification of individual frequencies results generally in the loss of sound naturalness, i. e., in sound distortions. Such distortions are referred to as amplitude-frequency distortions or in brief *frequency distortions*.

An amplitude-frequency (or simply frequency) response derived either by calculation or by measurement gives a pictorial indication of frequency distortions typical of an amplifier. The *frequency response* is the relationship between amplifier stage (or an amplifier as a whole) gain and frequency.

One of the possible frequency responses of an AF amplifier is illustrated in Fig. 9-2. To show more clearly the response behaviour near lower and upper limit frequencies, the logarithmic scale is used for frequencies, i. e., a non-uniform scale. The gain is given along the vertical axis either in absolute or relative units (reduced to the gain at a frequency of 400 Hz, for instance). The frequency response shown in Fig. 9-2 decays near the lower and upper limit frequencies

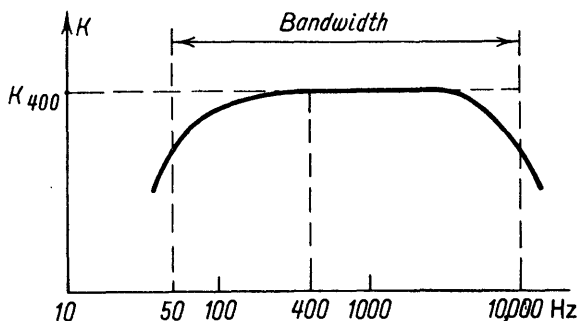


Fig. 9-2. Frequency response of audio-frequency amplifier (example)

as compared with middle frequencies. At some frequencies an increase in gain is also possible. One of the causes of a non-uniform frequency response are reactances inherent in electron valves (interelectrode capacitances), transistors (capacitances of $P-N$ junctions) or in external circuits.

It is impossible to provide a constant gain within a given frequency band. Fortunately, it is not required as a decrease in gain by 30 per cent (down to $0.7 K_{400}$, i. e., by 3 dB) hardly affects sound fidelity (or speech distinguishability in special communication systems). The band of frequencies located between the ordinates amounting to $0.7 K_{400}$ is called the *bandwidth*. Tone controls in modern receivers generally allow the bandwidth to be reduced both on the side of lower and upper frequencies in order to attenuate possible interference. The suppression of upper or lower frequencies makes the sound unnaturally dull or hissing.

The *sensitivity* of an amplifier, i. e., its ability to reproduce weak signals is increased with the increasing number of

amplifier stages but simultaneously the frequency non-uniformity is also increased. If, for instance, at the lower limit frequency the frequency response of a single stage falls to 0.7 as compared with its centre-frequency level, then three such stages lower the over-all frequency response to $0.7^3=0.34$ of the centre-frequency level. In other words, requirements for each stage are more rigid than for the amplifier as a whole.

The essence of frequency distortions consists in the fact that in the output signal the relationship between the components of various frequencies becomes quite different to those in the input signal. During amplification the relation between the phases of separate components in a signal may also change since phase shifts depend upon frequency. Such distortions are generally called phase-frequency distortions or simply *phase distortions*. These distortions are of no importance for audio-signal fidelity as the human ear is insensitive to changes in relative phases of individual audio-signal components. In television, however, changes in relative phases of individual signal components result generally in image distortions and, therefore, they become an important factor.

Both frequency and phase distortions may be called linear since they are caused by stage capacitors and inductance coils, i. e., by elements which do not disturb the proportionality between voltage and current. As we know, the proper choice of initial voltages will cause an electron valve or transistor to operate on the linear portion of the characteristic curve (see, for instance, Figs. 8-22 and 8-51). Large amplitude signals, however, may run beyond the linear portion. In this case the output signal waveform will be distorted as shown in Fig. 9-3a. Such distortions are called *non-linear*. Non-linear distortions may also appear in the amplifier input circuit due to the grid current flowing in the circuit only with a positive voltage applied to the grid (see Fig. 8-28).

A periodic but non-sinusoidal signal, for instance, the anode current as shown in Fig. 9-3a, may be represented as a series of sinusoidal components having frequencies which are multiples of the fundamental frequency. These components are called the harmonics of the non-sinusoidal process. In Fig. 9-3b the upper waveform is a non-sinusoidal periodic signal and

the lower waveforms represent the first (fundamental) and second harmonics of this process. By adding the two harmonics algebraically, we obtain approximately the upper waveform. The distorted waveform of anode current in Fig. 9-3a contains many harmonics (theoretically an infinite number) and a direct component (i. e., a "zero frequency" harmonic).

Thus, mathematically *non-linear distortions* are considered as the *appearance of higher harmonics* not present in the original primary signal. In practice, these distortions deteriorate

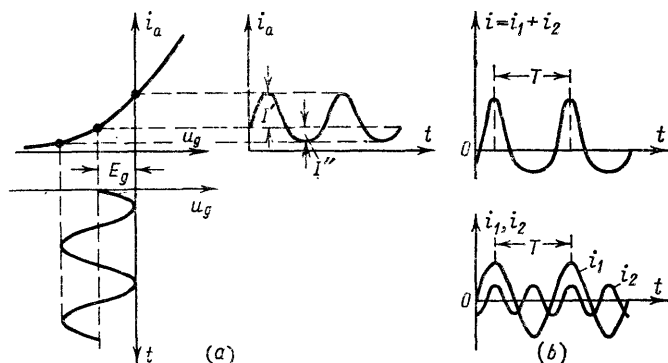


Fig. 9-3. Non-linear distortions

(a) distortion origin; (b) analysis of distorted signal

music harmony and blur speech. Therefore, the necessity for operation on the linear portion of the dynamic characteristic curve is one of the main requirements for an amplifier stage. It is more difficult to meet this requirement in the final stage than in the preceding ones since input amplitudes are greater at the final stage than at the preceding ones. So that the higher harmonics will not be detected by the ear, the amplitude of the most pronounced harmonic must not exceed 5 per cent of that of the fundamental harmonic. This ratio is used as the *non-linear distortion factor*.

The *dynamic range of the sound* reproduced after amplification, i. e., the difference between the levels of the most loud and most weak sounds, is also limited by the non-linear characteristics of active amplifying devices. Sometimes, even in this respect, requirements are imposed upon the amplifier

stage which is more sensitive to overloads than others, i. e., its operating point is more probable to run beyond the linear portion of the characteristic curve.

The operating point is the point on a characteristic curve corresponding to an instantaneous value of input voltage, i. e., the point which moves up and down along the operating portion of the characteristic curve as input voltage varies. At zero signal voltage the operating point is at the "initial" position (the quiescent point).

The presence of non-linear portions and specific characteristic curves of various amplifying devices force the use of a graphical method for determining the quiescent point (see Fig. 8-51). In case of large signals the relationships between voltages and currents in an active device are also non-linear (no proportionality exists between them). Therefore, to determine the alternating components of voltages and currents and, consequently, the amplification factor, graphical methods should be used. For small signals (about a fraction of a volt), the operating portion of the device characteristic curve may be considered linear. In this case, the alternating currents and voltages can be related through constant quantities such as transconductances. These constants may be called the *parameters* of an amplifying device for small signals.

In Chapter VIII we have already dealt with valve parameters and with some transistor parameters. They were determined from a family of characteristic curves. However, it will probably be more convenient to introduce the parameters which can be measured directly (so to say, externally) with the aid of conventional AC meters.

In fact, such "secondary" parameters which are common for both electron valves and transistors were suggested. Among these are the so-called *active transducer parameters*. In Fig. 9-1 an amplifier is shown in the form of a rectangle having two input terminals (one port) and two output terminals. This amplifier representation is called a transducer or two port. Because of the presence of active components (electron valves or transistors) the amplifier is called an active transducer. A unified system of parameters may be introduced for any active transducer and their values for each individual device may be tabulated in a manual so that one could use them in designing a required amplifier.

Let us assume *admittances* as common parameters and represent the amplifying device (valve or transistor) as a transducer (Fig. 9-4). The amplitudes of its input current $I_{1 \max}$ and output current $I_{2 \max}$ depend upon input voltage $V_{1 \max}$ and output voltage $V_{2 \max}$. Physically, it is clear that signal voltage $V_{1 \max}$ produces alternating currents at the input and output. However, both an electron valve and transistor have a feedback circuit formed by a grid-anode capacitance (see Fig. 8-29) or a collector-base conductance. Because of this circuit output voltage may affect the input current. The effect of both voltages on the input current may be expressed by Kirchoff's 1st law (sum of currents):

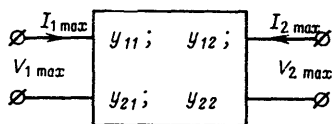


Fig. 9-4. Representation of valve or transistor as active transducer

$$I_{1 \max} = y_{11}V_{1 \max} + y_{12}V_{2 \max} \quad (9-8)$$

Here, admittance $y_{11} = \Delta i_1 / \Delta u_1$ fully determines input current when $V_{2 \max} = 0$, i. e., when output terminals are shorted (for alternating current). y_{11} is the *input admittance* of a device when its output is shorted. Incidentally, in Fig. 9-1 the output circuit will be shorted when $R_l = 0$. It can be easily shown that for a negatively-biased valve $y_{11} = j\omega(C_{gh} + C_{ag})$ and for a transistor $y_{11} = g_{11} + j\omega C_{11}$, with conductance g_{11} being dominant at lower frequencies. For a common-emitter circuit $R_{11} = 1/g_{11}$ generally amounts to about 1000 Ω .

Admittance $y_{12} = \Delta i_1 / \Delta u_2$ fully determines input current when $V_{1 \max} = 0$, i. e., when the input is shorted. With voltage $V_{2 \max}$ applied to the output terminals a current appears in the input circuit. y_{12} is the *feedback admittance* of a device when its input is shorted. Obviously, for an electron valve $y_{12} = j\omega C_{ag}$ is a pure capacitive susceptance. For a transistor the feedback admittance contains two components, $y_{12} = g_{12} + j\omega C_{12}$. Generally g_{12} is negligible in comparison with g_{11} and, therefore, at low frequencies we may assume $y_{12} \approx 0$, but at high frequencies we need only take into account feedback capacitive susceptance, $y_{12} = j\omega C_{12}$, where C_{12} amounts to a fraction of a picofarad or several picofarads (for the best common-emitter high-frequency transistors).

Directing our attention to output alternating current, we may also express its dependence upon both voltages by Kirchhoff's equation, namely,

$$I_{2 \max} = y_{21} V_{1 \max} + y_{22} V_{2 \max} \quad (9-9)$$

Here again, admittance y_{21} fully determines output current when $V_{2 \max} = 0$, i. e., when output terminals are shorted for alternating current. As we remember, in electron valves this *transfer admittance* $y_{21} = \Delta i_2 / \Delta u_1$ (for the output shorted) is called *transconductance* ($y_{21} = S$). This parameter characterizes valve *amplification properties* (see formula 8-15). In a common-emitter transistor, y_{21} is also an *amplification parameter* that can be shown as follows. As we know, parameter β is the current gain of a common-emitter transistor, when its output is short-circuited. Therefore, $I_{2 \max}$ (when $V_{2 \max} = 0$) may be expressed in the following form

$$I_{2 \max} = \beta I_{1 \max}; \quad I_{2 \max} = y_{21} V_{1 \max}$$

Consequently,

$$y_{21} = \beta \frac{I_{1 \max}}{V_{1 \max}} \approx \beta g_{11} = S$$

This parameter, as applied to a common-emitter transistor, we shall also call the *transconductance*. For transistors used in receivers it amounts practically to dozens and seldom hundreds of millisiemens (i. e., mA/V).

Finally, admittance $y_{22} = \Delta i_2 / \Delta u_2$ determines output current of a device when its input is shorted (i. e., when $V_{1 \max} = 0$). y_{22} is called the *output admittance* of the device. In electron valves operating at low frequencies the output admittance will be obviously the reciprocal of the anode (internal) resistance, $y_{22} = g_{22} = 1/R_i$, as follows from formula (8-9). At high frequencies the susceptance of anode-cathode capacitance may become appreciable and then $y_{22} = 1/R_i + j\omega C_{ak}$. For a transistor the meaning of the parameter y_{22} is just the same, i. e., $y_{22} = g_{22} + j\omega C_{22} = 1/R_{out} + j\omega C_{22}$. As indicated in Section 8-8, in a common-emitter transistor R_{out} amounts to approximately thousands of ohms, and the conductance is measured by tenths of a fraction of a millisiemens. This is quite a high conductance and at low frequencies it is much greater than ωC_{22} so that the latter can be neglected.

Thus, the above y -parameters of electron valves and transistors may be illustrated by the schematic diagrams shown in Fig. 9-5. Approximate values of the corresponding low-frequency parameters are tabulated in Table 9-1 (for illustration only).

It will be recalled once again that these values refer to the common-cathode and common-emitter connections of receiving amplifier devices. The values of y -parameters

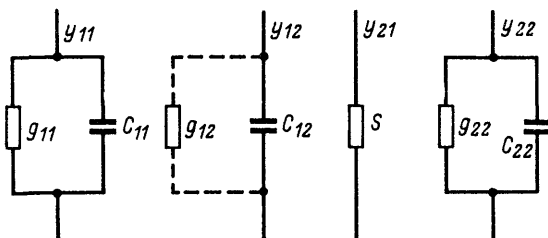


Fig. 9-5. Y -parameters of valve and transistor

change with increasing signal frequency. The main cause of this is the inertia of charge carriers. In electron valves, however, it becomes apparent at ultra-high frequencies, whereas in transistors it manifests itself at more lower frequencies

Table 9-1

Parameter	$R_{in} = 1/g_{11}$	C_{11}	g_{12}	C_{12}	S	$R_{out} = 1/g_{22}$	C_{22}
For electron valve	Up to ∞	Several picofarad	0	From 10^{-3} to 10^1 pF	From 10^1 to 10^2 mS	From 10^3 to 10^6 Ω	Several picofarad
For transistor	From 10^2 to 10^4 Ω	From 10^1 to 10^2 pF and more	From 10^1 to 10^2 μ S	From 10^1 to 10^2 pF and more	From 10^1 to 10^2 mS	From 10^3 to 10^4 Ω	From 10^1 to 10^2 pF and more

since the mobility of charge carriers in solids is less than in vacuum. Of particular importance for practice is the value of transconductance. For transistors it decreases with increasing

frequency according to the following law

$$S_f = \frac{S}{\sqrt{1 + \left(\frac{f}{f_{cutoff}}\right)^2}}$$

where S_f = transconductance at signal frequency f

S = low-frequency transconductance value

f_{cutoff} = cutoff frequency at which transconductance is decreased by $\sqrt{2}$ times (by 3 dB). The values of f_{cutoff} for each particular transistor are given in manuals. It is advisable to use transistors at $f \ll f_{cutoff}$

From the values given in the table one may conclude that electron valves cannot be replaced with transistors in a straightforward way since this requires the corresponding change of external components.

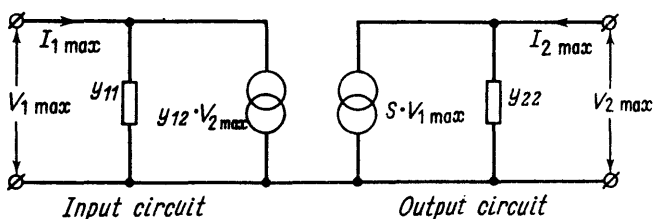


Fig. 9-6. Equivalent circuit of amplifier for the case of small signal

For small signal currents and voltages an amplifying device may be represented by the equivalent circuit shown in Fig. 9-6 which corresponds to equations (9-8) and (9-9). Signal voltage $V_1 \max$ is applied to the input terminals. It produces an amplified current $S V_1 \max$ in the output circuit. This current will branch off between the intrinsic admittance of the device, y_{22} , and the load admittance, thus producing voltage $V_2 \max$ across the output terminals. In turn, the input circuit presents for a signal, first, the admittance y_{11} and second, under the action of the output voltage, the feedback current, $y_{12} V_2 \max$, parallel to the input is induced in the circuit. This current may change the total current with a fixed signal level, i. e., speaking in terms of the equivalent circuit, it may change the *input admittance*.

We now turn our attention from the equivalent circuit of a device to the *equivalent circuit of an amplifier stage*, and consider first the *output circuit*. In Fig. 8-27 we have already encountered a schematic representation of an output circuit with a current source. A more general case of the same circuit is shown in Fig. 9-7a. Here again, some load admittance Y_l which gives the name to the stage (resistance amplifier,

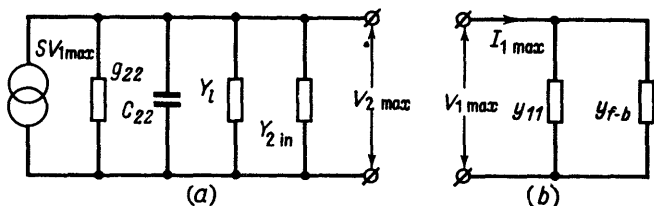


Fig. 9-7. Equivalent circuits of amplifier stage
(a) output circuit; (b) input circuit

choke-coupled amplifier, etc.) is connected into the output circuit. In addition, allowance is made for the fact that the output of a given stage is connected to the input of the next stage (final stages are not taken into consideration since signals in these stages cannot be assumed to be small). In Fig. 9-7a the succeeding stage is represented by its input admittance, $Y_{2\text{ in}}$.

From Ohm's law it follows that the output voltage is equal to the source current divided by the total admittance, i. e.,

$$V_{2\text{ max}} = \frac{SV_{1\text{ max}}}{Y_T}$$

where $Y_T = 1/Z_T$ is the modulus of the sum of admittances in all branches of the circuit under consideration. Hence, the *voltage gain*

$$K = \frac{V_{2\text{ max}}}{V_{1\text{ max}}} = \frac{S}{Y_T} = SZ_T \quad (9-10)$$

This equation is true for both electron valves and transistors. Subsequently we shall pay more attention to it.

We cannot, however, confine ourselves only to the analysis of a stage output circuit since in this case the input circuit also requires consideration for two reasons. First, it is ne-

cessary to know the total input admittance Y_{in} of the succeeding stage in order to compute the voltage gain since this admittance is included into the total admittance Y_T . Second, the input admittance of the given (being designed) stage must also be known so that the power gain can be calculated. The input circuit diagram shown in Fig. 9-6a allows the stage input admittance to be determined by dividing input current by input voltage. Considering formula (9-8), we obtain

$$\begin{aligned} Y_{in} &= \frac{I_{1 \max}}{V_{1 \max}} = y_{11} + y_{12} \frac{V_{2 \max}}{V_{1 \max}} = \\ &= y_{11} + y_{12}K = y_{11} + y_{f-b} \end{aligned} \quad (9-11)$$

Thus, the action of a feedback current source is equivalent to the connection of an additional admittance which we called the *feedback admittance*, $y_{f-b} = y_{12}K$ (Fig. 9-7b). For an electron valve this action was studied in detail for the first time in 1930 by V. I. Siforov.

For simplicity we will not give a general analysis of y_{f-b} , and consider some specific cases. Assume we have a valve amplifier stage whose output circuit is represented by a pure conductance $Y_T = G_T$ so that its voltage gain $K = S/G_T$. The input admittance of a valve with its output short-circuited is $y_{11} = j\omega(C_{gh} + C_{ag})$, and feedback admittance $y_{12} = j\omega C_{ag}$. Consequently, we may write

$$y_{in} = y_{11} + y_{12}K = j\omega(C_{gh} + C_{ag} + C_{ag}K)$$

Thus, the input admittance proved to be a pure capacitive susceptance, i. e., the action of feedback is equivalent to *an increase of stage input capacitance*, namely,

$$C_{in} = C_{gh} + C_{ag}(1 + K) \quad (9-12)$$

For a pentode the additional term $C_{ag}K$ is not generally large since capacitance C_{ag} itself is very small. For triodes, however, this additional term may be several times more than the value of $C_{gh} + C_{ag}$. This affects the reactive properties of the preceding stage.

We now consider a transistor amplifier whose output circuit is presented by a pure conductance. Its input admittance $y_{in} = y_{11} + y_{12}K = g_{11} + j\omega C_{11} + g_{12}K + j\omega C_{12}K$, i. e., both conductance and capacitive components included in the input

admittance have increased, namely,

$$\left. \begin{aligned} g_{in} &= g_{11} + g_{12}K \\ C_{in} &= C_{11} + C_{12}K \end{aligned} \right\} \quad (9-13)$$

But if $g_{12} \ll g_{11}$, then only input capacitance is noticeably increased due to intrinsic feedback.

Of general information on amplifiers still necessary to refer to the choice of initial operating conditions for amplifier stages. For valve amplifier stages some methods of voltage supply to the screen grid from an anode supply source have been illustrated in Fig. 8-32. The initial bias voltage for the first grid is also drawn from the anode supply source. The complete supply circuit for a

resistance-coupled valve amplifier is shown in Fig. 9-8. Besides load resistor R_L , the anode circuit includes a small resistor R_{fil} (about thousands of ohms) which is shunted by a large-capacitance capacitor C_{fil} (up to a microfarad). This is a *decoupling filter* whose function is to pass the anode current alternating component through capacitor C_{fil} directly to the cathode, thus with the

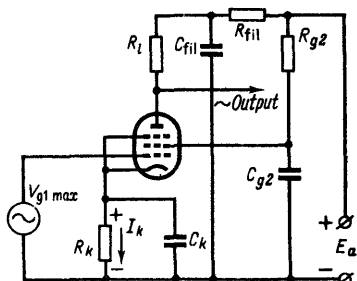


Fig. 9-8. Circuit connection of amplifier stage employing pentode

aid of resistor R_{fil} preventing the signal alternating current from passing through supply source E_a which is common for all amplifier stages ($E_a \approx 100$ to 200 V). In the absence of this filter the alternating voltage being produced by the current of the given stage across the internal resistance of the supply source would serve as a feedback voltage for the preceding stages causing possibly undesired changes in their admittances, thus deteriorating amplifier performance as a whole. For these reasons decoupling filters are generally used in multistage amplifiers. In the later schematic diagrams of individual stages we shall omit these filters but it is necessary to remember that they are required when connecting several amplifier stages together.

Bias voltage across the first grid, E_{g1} , is set up by cathode current I_k flowing from the top downwards (see Fig. 9-8)

through cathode resistor R_k (usually about a thousand ohms). The voltage drop

$$E_{g1} = I_k R_k = (I_a + I_{g2}) R_k$$

is *negative* with respect to the cathode. If it is necessary, for instance, to set bias $E_{g1} = 3$ V with cathode current $I_k = 2$ mA, the required resistance $R_k = E_{g1}/I_k = 1500$ ohms. Certainly, on the cathode resistor as well as on resistors in the anode circuit drops a small fraction of the supply voltage. The anode-to-cathode quiescent voltage remains equal to

$$V_{a0} = E_a - I_a R_l - I_a R_{fil} - I_k R_k \quad (9-14)$$

and voltage across the screen grid

$$V_{g2} = E_a - I_{g2} R_{g2} - I_k R_k$$

It is precisely V_{a0} , V_{g2} and E_{g1} values that determine the quiescent point position on valve characteristic curves.

Bias resistor R_k is shunted by large capacitor C_k (several microfarads). This capacitor must have a low (as compared with R_k) reactance $1/\omega C_k$ for a low-frequency signal so as to by-pass alternating current, thus, eliminating an alternating voltage drop across resistor R_k . It is easy to understand the necessity of such a capacitor if we take into consideration that an increase in signal positive voltage at the grid gives rise to cathode current, thus increasing grid negative bias. This means that the bias alternating voltage will be out of phase with signal voltage, i. e., will weaken useful input voltage. On one hand, large capacitors, C_k (electrolytic capacitors), used in each amplifier stage occupy considerable room in the entire amplifier. On the other hand, miniature solid versions of such capacitors are rather expensive.

In a transistor amplifier all stages are also supplied from a common source E_c (from 6 to 20 V). A simplified supply circuit for a transistor amplifier stage is shown in Fig. 9-9a. Here, R_l is a resistor presenting a specific case of a collector-circuit load while R_{fil} and C_{fil} are the components of a decoupling filter required for normal operation of multistage amplifiers (later we shall omit such filters in separate amplifier stages). Initial voltage (for *P-N-P* type transistors) is applied to the base, negative with respect to the emitter, through resistor R_1 connected between the base and the nega-

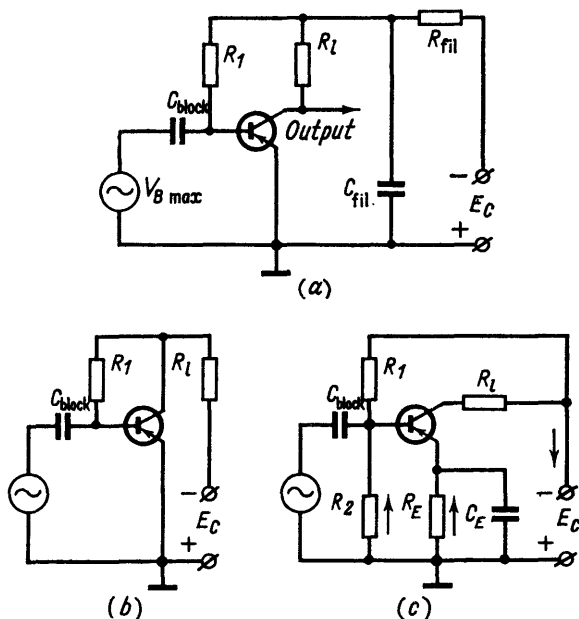


Fig. 9-9. Circuit connection of transistor amplifier stages
 (a) without temperature stabilization; (b) with collector stabilization; (c) with emitter stabilization and voltage divider

tive terminal of supply source E_C . On selecting the base direct current I_{B0} (see Fig. 8-51) and determining the appropriate bias V_{BE0} across the base, negative with respect to the emitter, we may find the resistance of the damping resistor R_1 from the equation *

$$R_1 = \frac{E_C - V_{BE0}}{I_{B0}} \approx \frac{E_C}{I_{B0}}$$

since $V_{BE0} \ll E_C$. The collector-to-emitter voltage is

$$V_{CE} \approx E_C - I_{C0}(R_L + R_{fil}) \quad (9-15)$$

Capacitor C_{block} isolates the signal source (i. e., the output of the preceding stage) from the action of DC voltage. In

* In some cases it is necessary to consider that the signal-source internal resistance is comparable with the transistor input resistance, therefore $V_{B \max} < E_{source \max}$.

transistor stages the reactance of this capacitor must appreciably be less than the stage input resistance even at low frequencies. Therefore, the capacitance of C_{block} exceeds sometimes dozens of microfarads.

This simple supply circuit, however, does not provide *stability* of the quiescent point and constancy of device parameters with *changing temperature*. Any increase in temperature disrupts the covalent bonds of an atom in a crystal and, consequently, increases the number of free charge carriers, thus increasing currents flowing through P - N type junctions. Temperature drift of characteristic curves is a disadvantage of transistors in general and common-emitter circuits in particular. Therefore, the circuit diagram shown in Fig. 9-9a is suitable only for transistors operating under constant temperature conditions (for instance, in domestic receivers).

Temperature stability is the problem of semiconductor technology. Temperature stability, however, may also be improved by a proper choice of the circuit configuration which is capable of maintaining constancy of operating conditions. Of course, circuit stabilization (or compensation) does not eliminate variations of transistor parameters with temperature.

The simplest method of stabilizing operating conditions in a common-emitter circuit is illustrated in Fig. 9-9b. This method is called *collector stabilization*. As compared with the previous circuit, here resistor R_1 is transferred from the supply source to the collector (the resistance, of course, has a new value). Physically, collector stabilization consists in the following. If current in the collector circuit increases with rising temperature, the DC voltage drop across resistor R_1 will also increase, and voltage V_C will drop. This, in turn, results in a drop of voltage across the base (which is negative with respect to the emitter), i. e., in a decrease of base direct current I_{B0} , thus leading to a decrease in collector current and, consequently, limiting the heating. Such an interaction is known as the *DC opposing coupling* (reverse transfer). Resistor R_1 connected between the collector and the base increases also AC feedback already existing due to the transistor intrinsic conductance g_{12} . This feedback, as we have seen, increases the circuit input conductance, i. e., weakens signal action.

In broadcast and television high-performance receivers, use is made of another supply and stabilization circuit which is shown in Fig. 9-9c. It is generally known as an *emitter stabilization* circuit employing a voltage divider, R_1R_2 , which is connected to the terminals of a supply source E_C to feed the base. The direct current flowing through the voltage divider must be 5 times as great as the initial base current selected for the transistor. Bias in the base circuit will then be approximately equal to the voltage drop across resistor R_2 and will remain essentially unchanged. Currents in the emitter and collector circuits, however, may change with temperature which is undesirable. Resistor R_E is connected in the emitter circuit to stabilize these currents. The function of resistor R_E is to produce between the base and the emitter an additional bias which acts *opposite* to the voltage developed across resistor R_2 . This is clearly seen from the direction of currents. If the emitter current rises, the voltage drop across resistor R_E will increase, and the negative base bias will decrease, thus resulting in a decrease of emitter and collector currents. To eliminate AC feedback, stabilizing resistor R_E is by-passed with a large-capacitance capacitor C_E (at least a dozen of microfarads).

The calculation of capacitor C_E is similar to that of capacitor C_h in the valve circuit shown in Fig. 9-8. From the stabilization viewpoint, it is desirable for resistor R_E to have an appreciable resistance, but it is limited by the permissible supply-voltage drop $\Delta V_E = I_{E0}R_E$ and, therefore, it does not exceed several kilohms. One more specific feature of this circuit must be considered: for input alternating voltage resistors R_1 and R_2 prove to be connected in parallel (to the common point) provided the circuit is connected to the preceding stage directly through blocking (coupling) capacitor C_{block} . Such a shunt increases the input conductance, which is undesirable. In the case of inductive coupling this shunting effect may be avoided as we shall see later.

9-3. AUDIO-FREQUENCY SMALL-SIGNAL AMPLIFIERS

While explaining the principles of amplification with the aid of electron valves and transistors (see Chapter VIII), we used circuits which included resistor R_l as the output load.

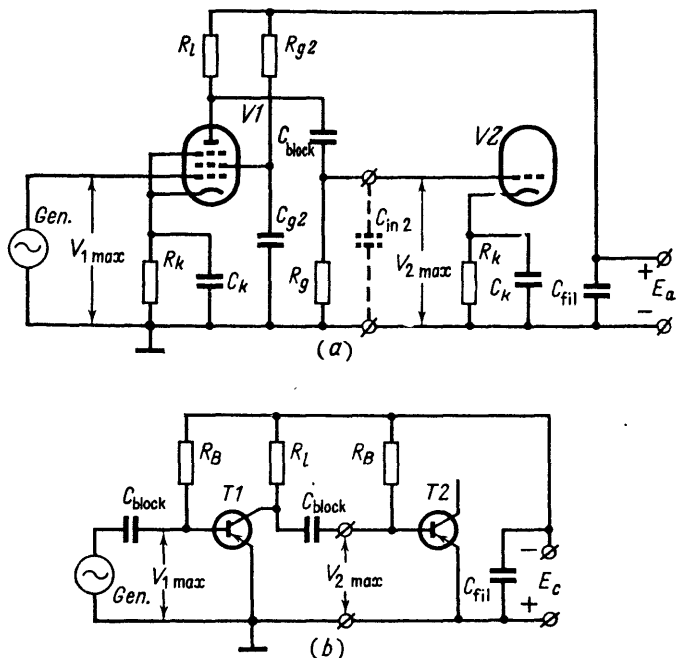


Fig. 9-10. Resistance-coupled amplifiers
(a) valve amplifier; (b) transistor amplifier

At the present time, resistance-coupled stages are extensively used for amplification of small signals and now we shall consider the operating principles of such stages in more detail.

Figs. 9-10 *a* and *b* illustrate valve and transistor resistance-coupled stages, respectively (in both circuits decoupling filters are represented only by capacitors C_{fil} which by-pass signal currents, thus preventing them from flowing through the supply source).

The valve stage is designed for signal *voltage amplification*. The amplified voltage is applied to the input of the succeeding stage to control its operation (the given and succeeding stages are built around valves $V1$ and $V2$, respectively). There is no need to mention here power amplification since grid circuits consume no power and the power delivered to resistor R_L is by itself not useful.

Valve V1 is a pentode although a triode may be used in a resistance-coupled stage as well. All the details concerning voltage supply and load are already known to us (refer to Fig. 9-8) except for capacitor C_{block} and resistor R_g . Input voltage $V_{1\max}$ is fed from a signal generator (*Gen.*) to the control grid of valve V1. Amplified voltage appears across resistor R_l connected into the anode circuit of valve V1. Blocking (coupling) capacitor C_{block} having a capacitance of dozens of thousands of picofarads is inserted between the output of the given stage and the input of the succeeding stage in order that the alternating voltage from resistor R_l act on the control grid of the succeeding valve as well as to isolate its control-grid circuit from the anode battery. This capacitor must offer high resistance for direct current. The presence of the blocking capacitor makes it necessary to connect resistor R_g between the grid and cathode of valve V2. This resistor serves two purposes. First, through this resistor a fixed negative bias voltage is applied from the lower terminal of resistor R_h to the control grid of valve V2. Second, resistor R_g provides a discharge path (leakage) for negative charges built up on the control grid of valve V2 due to electrons intercepted by it, thus preventing the valve from being cut off. Therefore, for this reason resistor R_g (hundreds of kilohms and more) is called the *grid-leak* resistor or simply the grid-leak.

That is how a preamplifier stage (voltage amplifier) employing resistors and an electron valve is made up. One should also take into account that the output terminals of the stage considered are shunted by input capacitor C_{in2} of the succeeding stage. Its capacitance is usually dozens of picofarads (included in this are grid-cathode capacitance, capacitance between connecting wires, and capacitance between amplifier components and its chassis). The output capacitance of valve V1 which shunts resistor R_l (see below) should also be included into capacitance C_{in2} .

The transistor amplifier (Fig. 9-10b) stage employs the common-emitter circuit connection of transistor T1. Voltage $V_{1\max}$ applied to the transistor base is fed from a signal generator (*Gen.*). Bias voltages are supplied to the bases of the given and succeeding transistors through resistors R_B . Load resistor R_l is connected into the collector circuit. AC

output voltage $V_{2\max}$ is taken from this resistor through blocking capacitor C_{block} . The input impedance of the succeeding transistor T_2 is a useful load of this stage. Here, we may speak about amplification of voltage, current and power.

An equivalent output circuit for valve and transistor stages with allowance for the input of a succeeding stage corresponds to that shown in Fig. 9-7a. The equivalent circuit, however, should be analyzed in detail before the selection of circuit parameters is discussed. A general equivalent circuit for

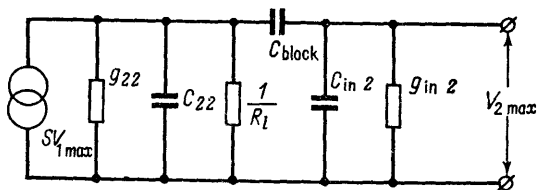


Fig. 9-11. Equivalent circuit of resistance-coupled amplifier

alternating currents and voltages is given in Fig. 9-11. It will be recalled that the conductance branch $1/R_l$ is connected in parallel with the output since for alternating current the upper end of resistor R_l is practically short-circuited to the common point through capacitor C_{fil} . It will also be recalled that in compliance with formula (9-10) *voltage gain* $K=S/Y_T$ proves to be *the greater, the less the total admittance* which is equal to the sum of admittances in all the branches.

Let us consider the selection of circuit parameters (component values) for a valve amplifier. Both a pentode and a triode may be employed in a resistance-coupled stage. For equal transconductances the pentode is more advantageous as its output conductance is less ($g_{22}=1/R_l$). On the other hand, the triode also has the advantage owing to linearity and uniform arrangement of its characteristic curves (see Fig. 8-19) as well as to its lower cost. As far as the input of the succeeding stage is concerned, in compliance with formula (9-12) capacitance $C_{in 2}=C_{gk 2}+C_{ag 2}(1+K_2)$, where $C_{gk 2}$, $C_{ag 2}$ and K_2 refer to the succeeding stage. $C_{in 2}$ amounts generally to several or dozens of picofarads. Valve input conductance g_{11} is zero, and therefore $G_{in 2}\approx 1/R_g$ where R_g

is a grid-leak whose resistance is about one megohm (see Fig. 9-10a).

The capacitance of blocking capacitor C_{block} , as stated above, must be sufficiently high. Even at moderate audio-frequencies (about 400-1000 Hz) its reactance $1/\omega C_{block}$ will be so small that capacitor C_{block} will present almost a short circuit. Therefore, the output capacitance of valve V_1 may be considered to be connected in parallel with the input capacitance of the succeeding stage so that the resultant capacitive branch will be $C_2 = C_{22} + C_{in\ 2}$. At moderate audio-frequencies (and all the more at lower) susceptance $\omega_m C_2$ is negligible as compared with conductances and, therefore, it is fair to consider capacitance C_2 equal to zero.

Thus, for moderate audio-frequencies the total admittance of the equivalent circuit shown in Fig. 9-11 is a pure conductance, i. e.,

$$Y_T = G_T = \frac{1}{R_i} + \frac{1}{R_l} + \frac{1}{R_g}$$

hence, voltage gain at these frequencies is

$$K_m = \frac{S}{G_T} \quad (9-16)$$

This voltage gain corresponds to the centre portion of the frequency response shown in (Fig. 9-2). In order to obtain higher gain for a given R_i and R_g , the resistance of R_l should be chosen as high as possible. However, as we know from formula (9-14), initial anode voltage V_{a0} decreases in this case, and the operating point may shift to a portion of the anode-grid characteristic having a lower transconductance (see, for instance, Fig. 8-24). As a result, gain K will turn to be lower rather than higher. For triodes load resistance R_l must satisfy the inequality $R_l \leq 3R_i$, i. e., it amounts to *thousands or dozens of thousands of ohms*. For pentodes load resistance R_l is taken of the same order, i. e., it proves to be much less than pentode (internal) resistance R_i . As we shall see later, circuit conductance should be not less than the shunt capacitive susceptance even for the upper frequency of the amplifier.

In the region of low audio-frequencies a part of the signal voltage developed by the alternating anode current across resistor R_l is wasted across the reactance of the blocking

capacitor so that output voltage decreases. This is illustrated in Fig. 9-2 by the decay of the frequency response in the low-frequency region. If the gain is allowed to fall at the lower frequency, ω_l , by 3 dB below its maximum level (i. e., up to 0.7 of the maximum gain), then the capacitance of blocking capacitor C_{block} should be determined from the equation

$$C_{block} \geq \frac{1}{\omega_l (R_l + R_g)} \quad (9-17)$$

As seen from Fig. 9-2, amplifier gain may also fall in the higher frequencies region. One of the reasons for this decay is the increase of susceptance of the resultant shunt capacitor C_2 whose value exceeds that of admittance Y_T , thus decreasing the gain according to formula (9-16). However, capacitance C_2 is determined by the electron valve used and cannot be changed. Therefore, it is necessary to check whether conductance G_T is high enough for the gain to be above a specified level at higher frequencies. Assuming a 3 dB decay at upper (cutoff) frequency ω_u , we must satisfy the following inequality

$$G_T \geq \omega_u C_2 \quad (9-18)$$

If this inequality is not met, the resistance of R_l should be reduced, thus sacrificing the gain value at center frequencies.

For a numerical example let us take a triode designed for amplification of audio-frequency signals, which has the following parameters: $S=2.5$ mS; $R_i=10^4 \Omega$. Recommended initial operating conditions are $V_{a0}=150$ V; $I_{a0} \leq 6$ mA; $E_g=-2$ V. The interelectrode capacitances are $C_{gh}=5$ pF; $C_{ag}=1.5$ pF; $C_{ah}=4$ pF. It is necessary to obtain gain $K \geq 15$ at center frequencies with 3 dB allowable decay at the lower frequency $\omega_l=600$ Hz and at the upper frequency $\omega_u=60,000$ Hz.

We begin our calculations with selection of a load resistance and set $R_l \approx 2R_i = 2 \times 10^4 \Omega$. The grid-leak resistance is set equal to $R_g = 5 \times 10^5 \Omega$. Then voltage gain at center frequencies can be computed from formula (9-16)

$$\begin{aligned} K = \frac{S}{G_T} &= \frac{S}{\frac{1}{R_i} + \frac{1}{R_l} + \frac{1}{R_g}} = \\ &= \frac{2.5 \times 10^{-3}}{0.152 \times 10^{-3}} \approx 16 \end{aligned}$$

which meets the requirement. The resistor in the cathode circuit must have resistance

$$R_k = \frac{E_g}{I_{a0}} = \frac{2}{6 \times 10^{-3}} \approx 330 \Omega$$

and anode supply voltage

$$E_a = V_{a0} + I_{a0} (R_l + R_k) = 150 + 6 \times 10^{-3} (2 \times 10^4 + 0.33 \times 10^3) \approx 272 \text{ V}$$

The blocking capacitor in the cathode circuit must have capacitance

$$C_k \geq \frac{10}{\omega_l R_k} = \frac{10^7}{600 \times 330} \approx 50 \mu\text{F}$$

The capacitance of blocking capacitor C_{block} is computed by formula (9-17)

$$C_{block} \geq \frac{1}{\omega_l (R_l + R_g)} = \frac{10^{12}}{600 \times 5.2 \times 10^5} \approx 3600 \text{ pF}$$

To check a gain decrease in the region of higher frequencies, we set the voltage gain of the succeeding stage, K_2 , to be equal to 16 and find that

$$\begin{aligned} \omega_u C_2 &\approx \omega_u [C_{ak} + C_{gk} + C_{ag} (1 + K_2)] \approx \\ &\approx 60,000 [4 + 5 + 1.5 (1 + 16)] 10^{-12} \approx 2 \mu\text{S} \end{aligned}$$

which is much less than the total conductance G_T equal to $152 \mu\text{S}$. Consequently, at higher frequencies the gain decay will be much less than the permissible one as evident from equation (9-18).

Power consumed by the anode circuit will be $P_0 = I_{a0} E_a = 6 \times 10^{-3} \times 272 \approx 1.6 \text{ W}$. A stage composed of the components having the calculated parameter values will meet the requirements specified above.

Referring to the transistor stage shown in Fig. 9-10b, let us consider the selection of its circuit parameters. For this circuit formulas (9-16) to (9-18) are also valid. The values of parameters, however, will be substantially different.

Let the selected transistor have the following parameters: $S = 30 \text{ mS}$; $g_{11} = 1 \text{ mS}$; $C_{11} = 200 \text{ pF}$; $f_{cutoff} = 3 \text{ MHz}$; $g_{22} = 0.1 \text{ mS}$; $C_{22} = 20 \text{ pF}$; $g_{12} = 0$; $C_{12} = 500 \text{ pF}$. Determine the circuit parameters of a resistance-coupled stage which provi-

des power gain $K_P \geq 200$ at an output current peak value $I_{2\max} \geq 2$ mA.

We begin with the choice of initial operating conditions. Collector-current DC component I_{C0} should exceed $I_{2\max}$ so that non-linear distortions can be avoided. Let $I_{C0} = 5$ mA. The initial collector-to-emitter voltage (in a common-emitter circuit) for low-power amplifiers is generally taken from 3 to 6 V. Assume that $V_{CE} = 6$ V. Supply source voltage E_C (see Fig. 9-10b) should be taken twice as high, i. e., $E_C \approx 2V_{CE}$ which in our case amounts to $E_C = 12$ V.

The excessive part of E_C is consumed by resistor R_I connected in the collector circuit. Consequently,

$$R_I = \frac{E_C}{2I_{C0}} = \frac{12}{2 \times 5 \times 10^{-3}} = 1200 \Omega$$

Base voltage V_{BE0} is chosen from transistor characteristics in such a way as to obtain V_{B0} which ensures current $I_{C0} = 5$ mA with $V_{CE} = 6$ V. Generally, voltage V_{BE0} is much less than E_C and, therefore, the resistance of damping resistor R_B is roughly taken to be equal to E_C/I_{B0} . In practice, the resistance of R_B amounts to dozens and hundreds of kilohms but it is still less than grid-leak resistance in valve circuits. Neglecting low conductances $1/R_B$ and g_{22} and setting $g_{in2} \approx g_{11} = 1$ mS, we find that for alternating current the collector circuit is terminated in resistors R_I and $R_{in2} = 1/g_{11}$ connected in parallel, thus

$$R_T \approx \frac{R_I R_{in2}}{R_I + R_{in2}} = \frac{1200 \times 1000}{1200 + 1000} \approx 550 \Omega$$

The AC dynamic characteristic curve will have a slope corresponding to the 550- Ω resistance and pass through point $I_{C0} = 5$ mA which is determined by a DC load line for $R_I = 1200 \Omega$.

Hence, voltage gain is

$$K = SR_T = 30 \times 10^{-3} \times 550 = 16.5$$

Static current gain (see Section 9-2) can be determined from the known transconductance and input conductance as

$$\beta \approx \frac{S}{g_{in}} = \frac{30 \times 10^{-3}}{1 \times 10^{-3}} = 30$$

But part of the alternating current from the source connected in the collector circuit is branched off through resistor R_l , thus not reaching useful load R_{in2} . This current is distributed proportionally to the conductances so that current gain is

$$K_c = \beta \frac{G_T - g_l}{G_T} = \beta \left(1 - \frac{R_T}{R_l} \right) = 30 \left(1 - \frac{550}{1200} \right) \approx 16$$

Power gain $K_P = K K_c \approx 250$, which meets the specified requirement. The peak value of alternating current in the base circuit will be $I_{B \max} = I_{2 \max} / K_c = 2 \times 10^{-3} / 16 \approx 125 \mu\text{A}$. Peak values of output and input voltages will equal 2V and 0.125 V, respectively.

If it is necessary to obtain a frequency response (for voltage gain) extending from $\omega_l = 600$ Hz to $\omega_u = 60,000$ Hz with a gain decay less than 3 dB, then the blocking capacitance determined by formula (9-12) will be

$$\begin{aligned} C_{block} &\geq \frac{1}{\omega_l (R_l + R_{in2})} = \\ &= \frac{10^6}{600 (1200 + 1000)} \approx 0.76 \mu\text{F} \end{aligned}$$

Note that this capacitance is much greater than that of capacitor C_{block} in a valve stage. Setting the voltage gain of the succeeding stage, K_2 , to be equal to about 16.5, we find that the total capacitance is $C_2 = C_{22} + C_{11} + C_{12}(1 + K_2) = 220 + 500(1 + 16.5) \approx 9000$ pF. Here, capacitances C_{11} and C_{12} were taken the same as for the 1st transistor. Let us determine the capacitive susceptance at the upper frequency from formula (9-18) and compare it with conductance $G_T = 1/550\text{S} = 1.82 \text{ mS}$:

$$\omega_u C_2 = 6 \times 10^4 \times 9000 \times 10^{-12} = 0.54 \text{ mS}$$

Consequently, at the upper frequency the gain decrease will be less than the permissible one. Here, we considered the fact that the transistor cutoff frequency, f_{cutoff} , was sufficiently high and allowed us to neglect the susceptance caused by a finite mobility (inertia) of charge carriers.

The purpose of this approximate calculation of a resistance-coupled amplifier employing a common-emitter transistor is to familiarize the reader with parameters which are typical

of transistor circuits and to show that stage operation in the presence of base current complicates substantially all calculations as compared with valve stage design.

If, distinct from the schematic diagram in Fig. 9-10b, the second stage employs a common-base circuit, the conditions for voltage amplification in the 1st stage are deteriorated considerably since the input resistance of a common-base circuit is very low (from 20 to 30 Ω). Satisfactory current gain can yet be obtained. If the amplifying properties of

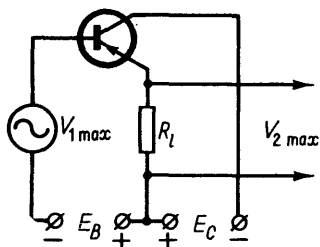


Fig. 9-12. Matching stage employing common-collector transistor

a common-emitter transistor cannot be fully realized due to connection of the high-resistance stage output to the low-resistance load, then a matching stage employing a *common-collector circuit* (Fig. 9-12) can be connected between them. In this configuration the signal source is inserted in the base circuit, the load (resistor R_L), in the emitter circuit and the collector is connected directly with the common point.

The base and collector circuits are supplied from different sources but it is more convenient to feed them from a single supply source since voltage signs across the base and collector are alike.

Without going into details of physical processes peculiar for the common-collector circuit (used comparatively seldom) we only say that it has a high input resistance (up to hundreds of thousands of ohms) and a low output resistance (dozens and hundreds of ohms). Consequently, this circuit can successfully be used as a matching stage. It provides current gain almost as high as β times owing to the fact that emitter current which in this circuit feeds load resistor R_L is always greater than base current. At the same time, the common-collector circuit weakens signal voltage, i. e., makes $K < 1$ due to the presence of feedback, i. e., the action of output current upon the input through resistor R_L . This feedback will be discussed in more detail in a special section.

This concludes our description of audio-frequency small-signal amplifiers. Circuits, other than resistance-coupled

ones, may be employed for these amplifiers such as of a transformer-coupled type but they are used primarily in output amplifiers and drivers which are considered in the next section. Resistance-coupled circuits will be encountered once again in wide-band amplifiers.

9-4. AUDIO-FREQUENCY OUTPUT AMPLIFIERS

In broadcast receivers and TV audio channels output (final) stages serve to deliver the required signal power to a loudspeaker or a set of loudspeakers. The power drawn by the output stages from the supply sources and their distortion levels should be as low as possible. To give some idea of the properties of AF power consumers, we shall say a few words about loudspeakers.

The operation of an *electrodynamic loudspeaker* is based on the interaction between a current-carrying conductor and a permanent magnet field. By its operating principle, this is the most extensively used type of a loudspeaker. Its magnetic system comprises a powerful permanent magnet with a circular air gap. The so-called voice coil is inserted into the air gap (an audio-frequency signal current is fed to the coil terminals from the amplifier output) and attached to a horn (made of an elastic compound). The horn may have the form of a circular or elliptical cone. The voice coil (together with the horn) can move freely in and out of the gap without touching the poles of the magnetic system. Such oscillations are performed by this elastic system because of the interaction between the permanent and alternating magnetic fields when alternating current flows through the turns of the voice coil. The function of the loudspeaker horn is to radiate sound waves in air.

The moving voice coil is the load for the amplifier output stage. With allowance made for conversion of electric power into acoustic energy the resistance of the voice coil generally amounts to a few ohms (rarely up to hundreds of ohms). Since the output resistance of an electron valve or transistor differs usually from the indicated value, the loudspeaker is connected to the output stage through a transformer providing a required load resistance. If a receiver has two loudspeakers (Sp_2 for higher audio frequencies and Sp_1 for lower

audio frequencies), they are connected to the secondary winding of a common transformer (Tr) through a high-pass filter (L_2C_2) and a low-pass filter (L_1C_1), respectively (Fig. 9-13).

Thus, we turn to the output stage coupled to its load resistor through a transformer. A schematic diagram of such a stage employing a beam tetrode is illustrated in Fig. 9-14a. The stage input is fed with alternating voltage $V_{1\max}$ supplied from the preceding stage.

The tetrode operates with no grid currents, which is achieved by selecting the correct cathode bias, i. e., the voltage drop across resistor R_k , which makes the control grid negative with respect to the cathode. Primary winding I of transformer Tr is connected in the anode circuit. Loudspeaker resistor R is inserted in the secondary winding II . Voltage is applied to the val-

Fig. 9-13. Connection of two loudspeakers for different frequency bands

ve screen grid directly from anode supply source E_a , which is permissible for output stages employing beam tetrodes or pentodes.

The stage simplified equivalent circuit for alternating components of anode current is illustrated in Fig. 9-14b. Here, R_l is the loudspeaker resistance "reduced" to the pri-

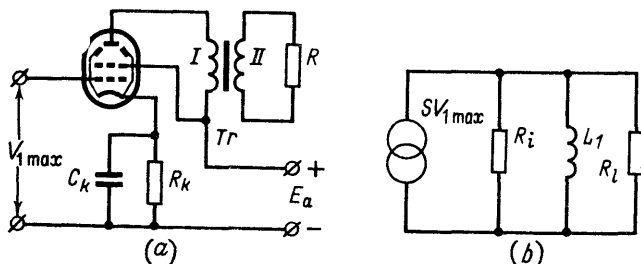


Fig. 9-14. Output stage employing beam tetrode with transformer-coupled loudspeaker

(a) schematic diagram; (b) anode-circuit equivalent network for alternating current

mary circuit of the transformer. This means that resistor R_l connected in the secondary winding is "brought" to the transformer primary circuit by finding a resistance value for R_l which would consume the same power as that delivered to the loudspeaker when resistor R_l is connected in the primary circuit. Voltages $V_{\max}$ and $V_{2 \max}$ across the primary and secondary windings are proportional to the number of their turns ω_1 and ω_2 , i. e., their ratio is equal to the transformation ratio n ,

$$\frac{V_{\max}}{V_{2 \max}} = \frac{\omega_1}{\omega_2} = n \quad (9-19)$$

whereas currents through the windings are inversely proportional to the number of turns

$$\frac{I_{\max}}{I_{2 \max}} = \frac{\omega_2}{\omega_1} = \frac{1}{n} \quad (9-20)$$

Consequently, the "reduced" value of load resistance is

$$R_l = \frac{V_{\max}}{I_{\max}} = \frac{V_{2 \max} n}{I_{2 \max} / n} = n^2 \frac{V_{2 \max}}{I_{2 \max}} = R n^2 \quad (9-21)$$

Inductance coil L_1 connected in parallel with resistor R_l is the inductance of the transformer primary (winding) when the secondary (winding) circuit is open (open-circuit inductance). The presence of inductive reactance $X = \omega L_1$ makes amplifier gain dependent upon the signal frequency since the useless current branched off in the inductance coil rises with the decreasing frequency. For the amplifier gain to be independent of frequency, the value of inductance L_1 should be selected such that its reactance will be greater than the "reduced" value of resistance R_l even at the lowest frequency f_l being amplified, i. e.,

$$2\pi f_l L_1 > R_l$$

If this inequality is satisfied, we may neglect the inductance in the schematic diagram (Fig. 9-14b) and consider that the valve is loaded only with resistor R_l .

When the lower frequency f_l (its rating has been specified in Section 9-1) and the resistance R_l are chosen, we may determine the desired inductance L_1 of the transformer primary. By selecting the volume of the transformer ferrite core and the diameter of the wire such that the losses in the

transformer will not exceed one or two dozen per cent of useful power, we may find the number of primary turns w_1 assuring the desired value of inductance L_1 (some thousands of turns of thin wire are generally required).

Then the desired transformation ratio n can be calculated from the reduced value of resistance R_l and the known value of the resistance R of the loudspeaker. The transformation ratio is determined from formula (9-21),

$$n = \sqrt{\frac{R_l}{R}} \quad (9-22)$$

As stated above, generally $R < R_l$, i. e., the transformer must step down voltage and step up current ($n > 1$). The number of secondary turns may be found from the equation

$$w_2 = \frac{w_1}{n} \quad (9-23)$$

The diameter of the secondary wire is determined by the current in the transformer secondary and may turn to be larger than the diameter of the primary wire.

Let us consider an example. Assume that the loudspeaker resistance $R = 10 \, \Omega$ and the desired anode load resistance $R_l = 2000 \, \Omega$. The lower frequency $f_l = 100 \, \text{Hz}$. Determine the basic parameter values for the transformer.

Assuming that at the lower frequency the reactance of the primary must be 1.5 times greater than the reduced resistance R_l , we find that

$$L_1 = \frac{1.5R_l}{6.28 \times 100} = \frac{3000}{628} \approx 4.8 \, \text{H}$$

Transformation ratio:

$$n = \sqrt{\frac{R_l}{R}} = \sqrt{\frac{2000}{10}} = 14.1 \text{ (voltage step-down ratio)}$$

This means that the number of secondary turns w_2 must be 14.1 times less than that of the primary turns.

What determines the choice of the value of anode load resistance R_l ? For each valve and each transistor under given operating conditions a most advantageous value of R_l exists, i. e., the value which provides the maximum useful power with a specified level of non-linear distortions. Let us explain

this by using the characteristic curves of a beam tetrode (or pentode).

Figure 9-15 shows a family of idealized anode characteristics of a valve for which the Manufacturer recommends an initial anode voltage of 300 V. Such a voltage should be delivered by the anode supply source since the DC resistance of the transformer primary is fairly low in comparison with the DC internal resistance of the valve; for direct current the load resistance R_L is zero.

When selecting the negative bias for the control grid, consideration should be given to the following.

First, in the absence of a signal the over-all supply power $P_0 = V_{a0} I_{a0}$ is dissipated on the anode. As known from Chapter VIII, the maximum anode dissipation value is specified for each valve. Moreover, this value determines the direct anode current.

Let the maximum anode dissipation for the valve used amounts to 6 W (see Fig. 9-15). Then, the direct anode current should not exceed

$$I_{a0 \max} = \frac{P_0}{V_{a0}} = \frac{6}{300} = 0.02 \text{ A} = 20 \text{ mA}$$

By drawing a perpendicular from point $V_a = 300 \text{ V}$ until it intercepts with a horizontal line corresponding to $I_{a0} = 20 \text{ mA}$, we find initial point m . It is located on the characteristic curve corresponding to bias $V_{g1} = -5 \text{ V}$.

Second, it is necessary that the valve operate with no grid current when a signal is applied to its control grid. From this it follows that the peak amplitude of signal voltage delivered by the preceding stage should never exceed the bias value (in this case $V_{1 \max} \leq 5 \text{ V}$). At the same time the negative peak amplitude when added to the negative bias produces an instantaneous grid voltage $V_{g1 \min} = -2 V_{1 \max}$ (in this case $V_{g1 \min} = -10 \text{ V}$).

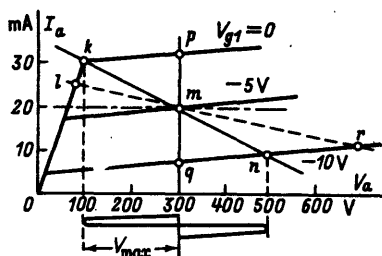


Fig. 9-15. Determining the optimum load resistance in anode circuit of output valve

Thus, let us assume the values determined by the selected position of point m to be the initial operating conditions. The dynamic characteristic whose slope is determined by the choice of the load resistance R_l should pass through this point. It will be recalled that when $R_l=0$ (the anode circuit is shorted), the signal will cause no changes in anode voltage so that it will remain equal to V_{a0} (in Fig. 9-15 $V_{a0}=300$ V). In this case the dynamic characteristic curve will be the vertical straight line pmq (as seen from the figure, the peak amplitude of alternating anode current will amount to approximately 12 mA). Useful power, however, is zero because of the absence of alternating voltage ($V_{\max}=0$) (in addition, load resistance is zero).

When the AC load resistance R_l is increased, the dynamic characteristic curve will assume an inclined position. In fact, an alternating voltage will be produced across R_l as anode current oscillates. With rising current the voltage drop across R_l increases and the anode voltage decreases and vice versa. The higher R_l , the smaller the slope of the dynamic characteristic curve.

For the optimum value of resistance R_l the dynamic characteristic will assume the position kmn in which segments km and mn are equal. In this case, the alternating anode voltage, $V_{\max}$, will be appreciably large. The useful power developed across R_l , i. e., delivered actually to the loudspeaker through the transformer, will be

$$P = \frac{V_{\max} I_{\max}}{2} \quad (9-24)$$

where $I_{\max}$ is the peak (maximum) amplitude of alternating anode current.

For this case, we neglect power losses in the transformer (due to heating of wires and the core) since these losses constitute a relatively small part of the useful power. In our example $I_{\max} \approx 10$ mA and $V_{\max} \approx 200$ V. Hence,

$$P \approx \frac{10 \times 10^{-3} \times 2 \times 10^2}{2} = 1 \text{ W}$$

This power is quite sufficient to provide a normal sound level from a domestic receiver's loudspeaker.

When the value of the resistance R_l corresponds to the position kmn of the dynamic characteristic curve, segments km and mn are equal and, therefore, anode current changes up and down by the same value, i. e., the electron valve does not produce any non-linear distortion (which we observed in Fig. 9-3).

If the value of resistance R_l is further increased, the dynamic characteristic curve will assume the position lmr . In this case anode current oscillations will decrease and, what is worse, they will be asymmetrical, i. e., non-linear distortions will appear. As a result, useful power will decrease.

The optimum value of load resistance for electron valves is generally about thousands or several dozen of thousands of ohms. In some cases, this value is specified by the Manufacturer in the valve certificate. It may also be determined, if necessary, by the above technique. In our example the optimum value of load resistance is found from the equation

$$R_l = \frac{V_{\max}}{I_{\max}} = \frac{200}{10 \times 10^{-3}} = 20,000 \Omega$$

The above operating conditions under which anode current flows through the valve at all time, varying symmetrically with respect to DC component I_{a0} is called class A amplifier operation. The major disadvantage of this mode is low efficiency η , i. e., a comparatively low output power with a relatively high power drawn from the anode supply source. The efficiency is defined as

$$\eta = \frac{P}{P_0} = \frac{V_{\max} I_{\max}}{2V_{a0} I_{a0}} \quad (9-25)$$

In the example discussed above

$$\eta = \frac{200 \times 10 \times 10^{-3}}{2 \times 300 \times 20 \times 10^{-3}} = \frac{1}{6} \approx 0.166 \text{ (about 17 per cent)}$$

It should be noted that this efficiency value corresponds to the maximum signal level. With no audio signal present, efficiency is zero and power consumption is also zero.

Transistors can also be employed in the output stage of an amplifier. It should be noted, however, that transistors designed for power amplifiers require a special finned radiator to increase heat-removal. Without the radiator the transistor

may heat up to an impermissible temperature even with a small power dissipated in its P - N junctions.

In output stages either the common-base or common-emitter circuit configurations are used. The common-base circuit is of a greater advantage since the power required for the loudspeaker is produced with a lower non-linear distortion level. This may be explained by the fact that the input and output characteristic curves of the common-base circuit are

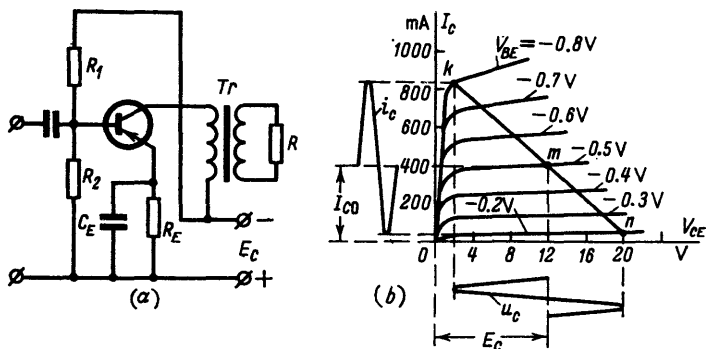


Fig. 9-16. Output stage employing the common-emitter transistor
(a) schematic diagram; (b) collector-circuit operation

more linear (see Figs. 8-49 and 8-51). On the other hand, the common-emitter circuit requires smaller input signal levels to provide the same output power. We shall dwell here upon the common-emitter circuit as it is the one most commonly used in the "push-pull" version which we will discuss later. This version reduces non-linear distortion at the output since higher harmonics cancel each other in this circuit.

A schematic diagram of a common-emitter output stage employing emitter stabilization and a transformer output is given in Fig. 9-16a. This diagram is self-explanatory. Fig. 9-16b shows the output characteristic curves of a specific power transistor. These characteristic curves as compared with the characteristics shown in Fig. 8-51 were taken for fixed base voltages (with respect to the common point) rather than for fixed base currents. Basically, however, there is no difference between them.

Let us determine the initial operating point m on the family of characteristic curves. Assuming that the supply voltage $E_C = 12$ V and neglecting (with some error) the DC resistance of the transformer primary, we search for point m on the vertical line corresponding to $E_C = 12$ V (certainly, this voltage is negative for a $P-N-P$ transistor). If the maximum collector dissipation power for the transistor used amounts, for instance, to $P_{C \max} = 6$ W, we must limit the collector current to $I_{C0} \leq P_{C \max} / E_C = 6/12 = 0.5$ A. Setting current I_{C0} to be equal to 0.4 A, we determine the position of point m and, consequently, the required supply power for the stage

$$P_0 = I_{C0} E_C = 0.4 \times 12 = 4.8 \text{ W}$$

Let us now determine the suitable resistance value for the load resistor connected in the collector circuit. For this purpose, we draw the dynamic characteristic curve through point m , i. e., an inclined line which has its lower end on the characteristic curve corresponding to the minimum (quiescent) collector current (point n) and its upper end on the bending point of a static characteristic curve (point k). If possible, segments km and mn should be equal to each other. We may assert that when the base voltage varies between -0.8 and -0.2 V, i. e., when input voltage $V_{1 \max} = 0.3$ V, the output stage is used effectively. Very approximately useful power equals:

$$P = \frac{V_{C \max} I_{C \max}}{2} \approx \frac{9 \times 400 \times 10^{-3}}{2} = 1.8 \text{ W}^*$$

Inequality of segments km and mn (caused by the lack of parallelism of actual static characteristics) leads to non-linear distortions which may be observed in the collector current i_C and collector voltage u_C plotted as a function of time.

In the given example the optimum value of load resistance R_L amounts only to about 20Ω . Thus, the transformation ratio will not differ appreciably from unity. Furthermore, sometimes it proves possible to connect an electrodynamic loudspeaker in the collector circuit without a transformer. For power transistors low-resistance loads are typical since a high useful

* The numbers of 9 V and 400 mA are obtained as the half-sums of voltage and current negative and positive half-waves approximately equal to the peak amplitudes of first harmonics.

power is provided by them with low collector voltages at the expense of high currents.

The above types of class *A* output amplifiers may be called single-ended amplifiers as distinct from push-pull amplifiers which will be discussed later. The power which a valve or transistor is capable to provide with a given distortion level may not be sufficient. To retain the type of electronic device, two valves or two transistors may be connected in parallel to each other. In this case, circuit parameters values will be, of course, different but we shall have to provide class *A*

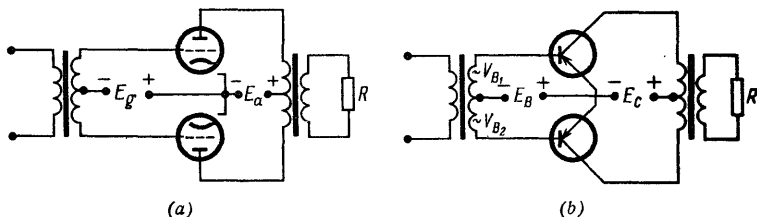


Fig. 9-17. Output-stage push-pull circuits

(a) valve (triode) stage; (b) output stage employing common-emitter transistors

operation as before. When using two valves or two transistors in one stage, it is more advantageous to connect them in the push-pull circuit as in this case distortions may be cancelled out.

Push-pull output stages employing valves (in the given case triodes) and transistors are shown in Fig. 9-17. As the operating principles of both circuits are the same, we shall consider the processes occurring in the transistor circuit only (Fig. 9-17b).

A push-pull stage is basically a combination of two single-ended stages which form two circuit "arms" having a common load R and working in phase opposition. At its input the push-pull stage is connected to a transformer having a centre tap from the secondary winding whereas at the output the push-pull stage has a transformer whose primary winding is also centre-tapped.

Direct currents in the collector circuits produce opposing ampere-turns in the halves of the output transformer primary and, therefore, the transformer core is not magnetized by

direct current. This permits the core volume to be reduced. Moreover, in such circuit configuration the results of collector current changes caused in the two halves by a poor supply circuit filtration cancel each other, whereas in the case of a single-ended stage they produce hum in the loudspeaker.

Assume that a signal alternating voltage is applied to the stage input. The centre tap of the input transformer secondary is connected to the circuit common point. Voltages fed to the bases from the secondary ends are opposite in sign (i. e., are displaced in phase by 180°). Consequently, when voltage across one of the bases (u_{B1}) rises (becomes more negative), voltage (u_{B2}) across the other base decreases and vice versa (Fig. 9-18 *a* and *b*). Collector current (only AC components are shown in the figure) in the first transistor (i_{C1}) rises accordingly, whereas in the second transistor collector current (i_{C2}) decreases and vice versa (Fig. 9-18 *c* and *d*). In other words, in the circuit portion which is common for collector and emitter circuits, i. e., in the supply source, no alternating currents are present as they cancel each other.

In the two halves of the output transformer primary the AC components of collector currents are not only shifted in phase by 180° but they flow from the centre tap in opposite directions. This means that the alternating magnetic fluxes (am-

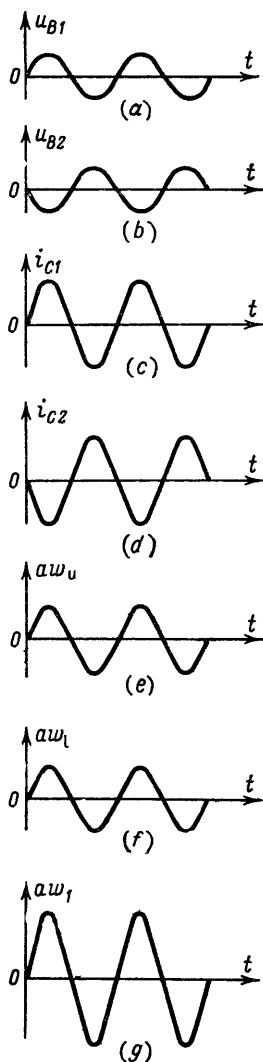


Fig. 9-18. Processes in push-pull stage

pere-turns aw_{upper} and aw_{lower}) produced by the AC components of collector currents in the upper and lower halves of the primary are in phase (Fig. 9-18e and f), so that the resultant flow aw_1 is doubled (Fig. 9-18g). EMFs induced in the secondary of the transformer by both transistors are in the same direction, and so twice the useful power can be obtained provided the arms are symmetrical.

This is precisely how the push-pull circuit functions. However, power could be doubled by connecting the valves in parallel. One remarkable property of a push-pull circuit is that non-linear distortions appearing in two valves cancel each other. Let us return to Fig. 9-3. As seen from the figure, when operating on the non-linear portion of the characteristic curve, one of the half-waves is "compressed". It is also shown that the AC component of distorted current contained not only a fundamental harmonic but higher harmonics as well (in particular, the second). Suppose now that such distortions occur in both transistors (or valves) of a push-pull circuit and both distorted currents induce magnetic fluxes in the core of the output transformer. The following is remarkable. When the first transistor produces a "sharp" current half-wave, the second one produces a "flat" (compressed) half-wave and vice versa (Fig. 9-19a and b).

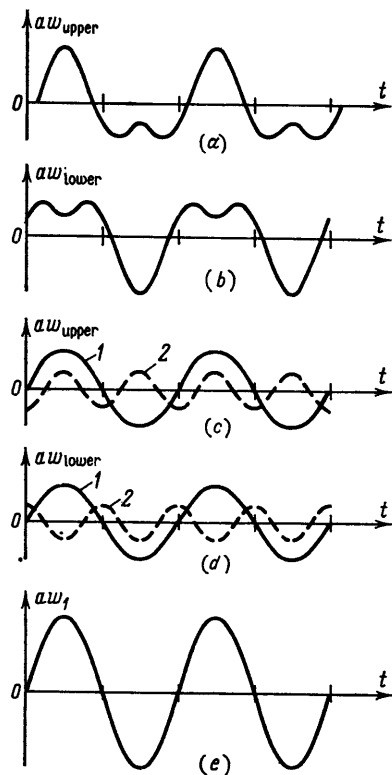


Fig. 9-19. Cancellation of 2nd harmonics in push-pull circuit

Decomposing the magnetic fluxes (ampere-turns) produced

by the upper (ω_{upper}) and lower (ω_{lower}) halves of the winding into the first and second harmonics and neglecting the fourth, sixth and other harmonics (Fig. 9-19c and d), we see that their fundamental harmonics (1) are in phase (as may be seen from Fig. 9-18) and that the second harmonics (2) are out of phase. Consequently, no second harmonics will be present in the resultant magnetic flux as they cancel each other, and only the fundamental-frequency flux (ω_1) of double amplitude (Fig. 9-19e) will exist in the transformer core. The fourth, sixth harmonics, etc. will be cancelled out in a similar way.

From this remarkable property it follows that for each arm of a push-pull circuit point m (see Figs. 9-15 and 9-16) may be selected not in the middle of a straight line kmn but somewhat below so that when at rest the current direct component is reduced, thus increasing useful power and efficiency of the stage. Higher harmonics (the second, fourth, etc.) appearing in each arm are cancelled out. Possible are operating conditions, when each arm of the push-pull circuit produces a current pulse only during a half-cycle, whereas no current flows through the collector (or anode) circuit during the next half-cycle. This mode is referred to as class B operation. An intermediate mode similar to that shown in Fig. 9-3 and known as class AB operation is often employed in amplifiers.

9-5. DRIVER STAGES. FEEDBACK IN AMPLIFIERS

A transformer (see Fig. 9-17) having a centre-tapped secondary winding is generally used to couple a single-ended driver stage to a push-pull output stage. A transformer coupling, however, may also be used for driving a single-ended output stage, especially in a transistor amplifier having an output stage with a low input resistance (for instance, a common-base transistor). A schematic diagram of a driver stage employing a common-emitter circuit with transformer Tr connected in the collector circuit is illustrated in Fig. 9-20a. The centre-tapped secondary winding of the transformer is connected to the input of a power push-pull stage employing two common-base transistors. The transformer-coupled stage is loaded with resistor R and capacitor C which are inserted between the emitters of the output stage's transistors.

In the equivalent circuit of the output circuit (Fig. 9-20b) the "reduced" values of $R_l = Rn^2$ and $C_l = C/n^2$, where $n = \omega_1/\omega_2$ is the transformation ratio, are shown.

One should take into account that part of the transformer magnetic flux does not participate in the energy transfer from the primary to secondary circuit as the lines of force

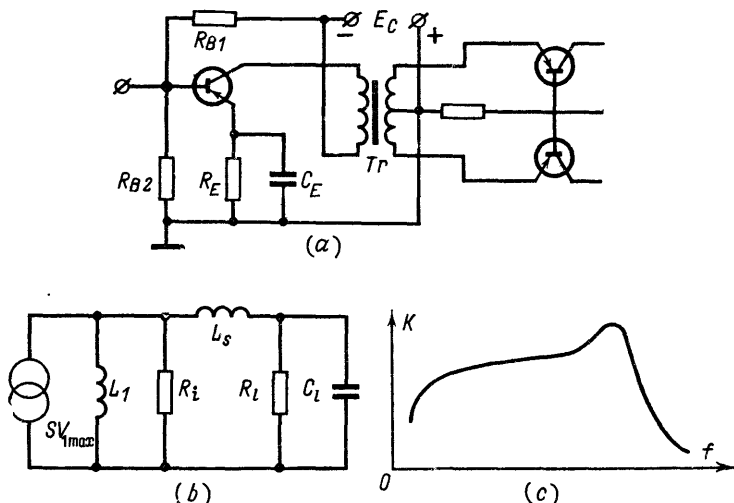


Fig. 9-20. Transformer-coupled driver stage

(a) schematic diagram; (b) output-circuit equivalent; (c) frequency response of valve driver

of the primary flux spread outward from the winding and so they do not thread through the secondary winding turns. This part of the alternating magnetic flux produces stray inductance L_s which is series connected in the equivalent circuit. In the region of higher audio frequencies this inductance may have an appreciable reactance ωL_s .

Suppose that we use a properly designed transformer, i. e., the one which has a sufficiently large primary "no-load" inductance L_1 , the fairly small stray inductance L_s and which allows for the susceptance of the "reduced" capacitance ($\omega C/n^2 \ll 1/Rn^2$) to be neglected over the entire band of audio frequencies. In this case, the equivalent circuit will include a current source $SV_{\max}$ operating on parallel-connected resis-

tors R_i and R_l . The function of the transformer in this stage consists in *matching* the load resistance to the internal one. If $R_l = R_i$, then *maximum power* will be delivered to the load (in compliance with a well-known electrical engineering theorem). Consequently, the optimum transformation ratio will be

$$n_{opt} = \sqrt{\frac{R_i}{R}} \quad (9-26)$$

Thus, the transformer-coupled stage not only makes it possible to couple a single-ended amplifier to a push-pull one but it matches its high-resistance output to a low-resistance input of the succeeding stage.

As distinct from the transistor stage, a valve transformer-coupled stage requires a rather high load resistance if the succeeding stage operates without grid current. Therefore, at upper frequencies one cannot neglect the effect of capacitor C_l , and stray inductance L_s connected in series with the capacitor cannot be neglected either. The fact is that series resonance may occur in circuit $L_s C_l$ in the upper frequency region. As is known from Chapter II, at such a resonance voltage increases across each of these reactive elements. Under these conditions the frequency response of the transformer-coupled stage may have a resonance peak at upper frequencies (Fig. 9-20c) since output voltage is taken from the capacitor's terminals. This property of the transformer-coupled stage may be useful if in other stages gain in the region of upper frequencies drops too drastically. This is a technique used for "flattening" (equalization) the over-all frequency response curve whose ordinates are the products of corresponding (in frequency) ordinates of the individual stage frequency responses. It should be noted that an excessive stray inductance in a transistor-coupled stage may lead to a gain decrease at upper frequencies but resonance phenomena do not occur in it due to a high load conductance.

Let us briefly dwell upon the notion of *feedback* in audio-frequency amplifiers. As stated earlier, feedback is the return of a portion of the output signal of an amplifier to its input. As is known, a feedback path may be a valve anode-(control) grid capacitance or a common portion of the crystal in transistor circuits. In the region of audio frequencies the coupling

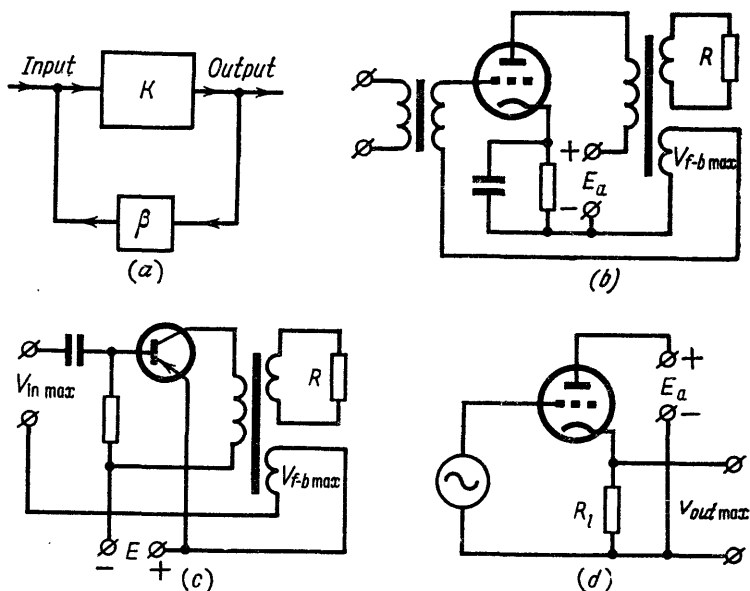


Fig. 9-21. Negative feedback

(a) feedback loop; (b) output stage employing a triode with negative feedback; (c) output stage employing a transistor with negative feedback; (d) common-anode stage

through an interelectrode capacitance (especially for valves with a screen grid) does not usually affect the amplifier performance. The effect of intrinsic feedback in a transistor is partially incorporated in the shape of its static characteristic curves. The reduction of stray coupling (including the interelectrode coupling) presents a more serious problem in high-frequency amplifiers.

Feedback most frequently used in audio-frequency amplifiers is introduced deliberately to improve some of their properties with a feedback signal being 180° out of phase with respect to the input signal. This deliberately introduced feedback is referred to as the *negative feedback*. A block-diagram of a general feedback circuit is illustrated in Fig. 9-21a. A feedback path along with an amplifier circuit to which it is connected forms a closed circuit known as a feedback loop. In general, the feedback loop may include one or several stages.

Let us estimate the effect of feedback on the voltage gain of an amplifier. Let the amplifier without a feedback path provides the voltage gain $K=V_{2 \max}/V_{1 \max}$. In the same amplifier fitted with a feedback path a portion of its output voltage is fed back to the input, and this portion amounts to $V_{f-b \max}=\beta V_{2 \max}$ where β is a nondimensional feedback ratio indicating which fraction of output voltage is fed back to the input through the feedback path. For negative feedback, β is usually ranges from 0.05 to 0.2.

In the absence of feedback input voltage $V_{1 \max}=V_{s \max}$ where $V_{s \max}$ is the voltage produced by the signal only. When negative feedback is present,

$$\begin{aligned} V_{1 \max} &= V_{s \max} - V_{f-b \max} = \\ &= V_{s \max} - \beta V_{2 \max} = V_{s \max} - \beta K V_{1 \max} \end{aligned}$$

Combining the terms with $V_{1 \max}$ and bringing $V_{1 \max}$ out of the brackets, we find that $V_{1 \max}=V_{s \max}(1+\beta K)$. Consequently, in the presence of negative feedback the voltage gain, i. e., the ratio of output voltage to *input signal voltage only*, is

$$K_{f-b} = \frac{V_{2 \max}}{V_{s \max}} = \frac{V_{2 \max}}{V_{1 \max}(1+\beta K)} = \frac{K}{1+\beta K} \quad (9-27)$$

Assume, for instance, that the gain of a stage without feedback is $K=20$ and the negative feedback ratio $\beta=0.1$. Then in the presence of feedback the gain will be

$$K_{f-b} = \frac{20}{1+0.1 \times 20} \approx 6.7$$

Consequently, negative feedback decreases the gain of a stage (or several stages) by $(1+\beta K)$ times. It may seem that amplifier performance worsens due to negative feedback. In fact, it is necessary to increase input voltage by $(1+\beta K)$ times in order to retain the same power which was delivered to the output in the absence of feedback. This means, however, that amplifier input resistance has increased by the same number of times so that an increased signal voltage can be obtained from the preceding stages.

What are then the advantages of applying negative feedback?

First, it *reduces non-linear distortion*. In fact, the higher harmonics of the signal frequency produced at the stage output are fed back to the input through the feedback path and after amplification they appear at the output being shifted in phase by 180° with respect to the harmonics which were caused by non-linearities of the stage. This means that higher harmonics are reduced, whereas the fundamental-frequency signal retains its output amplitude owing to the increased input voltage. The ratio of the peak amplitude of an undesired harmonic (for instance, the second) to the peak amplitude of the fundamental-frequency signal is reduced by $(1+\beta K)$ times.

Second, negative feedback *provides a more uniform frequency response* of an amplifier. This can be easily shown by the following example. Assume that in the absence of feedback the stage gain at centre and lower frequencies is $K_c = 20$ and $K_l = 10$, respectively. Thus, a highly non-uniform frequency response results in irregularity $K_c/K_l = 2$. On introducing feedback with $\beta = 0.1$, gain will be $K_{c, f-b} = 20/(1+0.1 \times 20) \approx 6.7$ and $K_{l, f-b} = 10/(1+0.1 \times 10) = 5$. As we see the irregularity will be much less, precisely $K_{c, f-b}/K_{l, f-b} = 6.7/5 = 1.3$. It should be noted, however, that the required phase shift (180°) of feedback voltage may be disturbed over the amplifier passband.

Third, negative feedback *improves stability of the gain* of an amplifier stage. Gain variations are caused primarily by changes in valve parameters and particularly in transistor parameters when a valve or transistor is replaced as well as by temperature variations, and by other factors. It is best to illustrate the stabilization property by considering an extreme case. Let us assume that the feedback ratio is high enough so that $\beta K \gg 1$. Then

$$K_{f-b} = \frac{K}{1+\beta K} \approx \frac{1}{\beta}$$

i. e., gain depends only upon the parameters of the feedback circuit rather than on valve or transistor parameters. In turn, the parameters of the feedback circuit are comparatively stable. The physical concept of stabilization is such that in the case of worsening parameters (for example, transconductance) the output voltage corresponding to a given input

signal falls together with the feedback voltage which is in antiphase to the signal. This means that the output signal is weakened less than it would be without feedback.

One could still indicate other positive results from the application of negative feedback but the advantages listed above are quite sufficient to prove its use.

Circuits employing a triode and transistor (see Fig. 9-21 *b* and *c*) are the simplest examples of output stages with negative feedback. In these circuits feedback voltage is produced by the third winding of the output transformer.

Feedback circuits having $\beta=1$ hold a somewhat specific position. Among these are the well-known common-collector circuit (see Fig. 9-12) and the common-anode valve circuit which is analogous to the former (Fig. 9-21*d*). In such circuits load resistor R_l is connected in the cathode circuit and the emitter circuit, respectively. The voltage produced by an alternating cathode (or emitter) current across this resistor constitutes the output voltage. But this very resistor is also included in the input circuit. This means that the output voltage is fully applied to the input in series with signal voltage, i. e.,

$$\beta = \frac{V_{out\ max}}{V_{f-b\ max}} = 1$$

Thus, stage gain $K_{f-b} = \frac{K}{1+K} < 1$, i. e., the stage does not amplify voltage. However, remarkable here is the fact that the input resistance of this circuit is extremely high, and the output resistance is low. As a result, this circuit is generally used as a matching stage. Also remarkable is the fact that, distinct from other circuits, the output voltage of the present circuit is in phase with the input one. This can be easily seen by considering the result of an increase in input voltage. Owing to the last property the transistor circuit of this type is called the emitter follower and the valve circuit is termed the cathode follower.

9-6. VIDEO AMPLIFIERS

While classifying primary-signal amplifiers in Section 9-1, we subdivided them into audio and video amplifiers emphasizing the fact that the latter are generally known as wideband

amplifiers. We have now to explain what dictates a wide bandwidth of video amplifiers and how it can be realized.

For amplifying video pulses and audio-frequency signals a valve should operate without grid currents. A resistance-coupled stage (see Fig. 9-10a) is the most advantageous type of the amplifier stage.

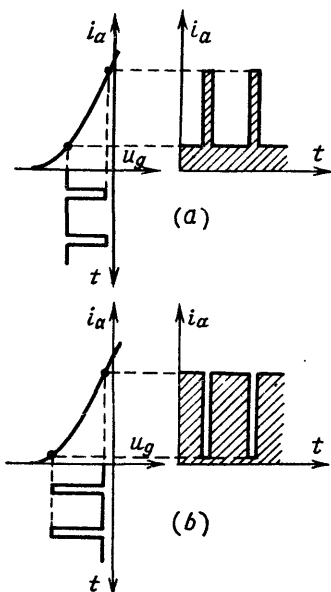


Fig. 9-22. Video-pulse amplification

(a) positive pulses; (b) negative pulses

In case of large pulses the operating point is selected in the lower or upper portions of the valve characteristic curve rather than in the middle of its linear portion. If positive pulses are applied to the control grid, the operating point is selected in the lower portion of the characteristic curve (Fig. 9-22a). These positive pulses cause anode current to rise and anode voltage to drop. If negative pulses are applied to the grid (Fig. 9-22b), the operating point is selected in the upper portion of the characteristic curve as shown in Fig. 9-22b. Negative pulses decrease anode current so that anode voltage rises. Thus, output voltage has an opposite sign (phase) with respect to the input one.

As we know, phase inversion also occurs during amplification of sinusoidal oscillations—with the increasing grid voltage anode current grows and anode voltage drops, and vice versa (see, for instance Fig. 8-22). It is precisely from the anode-cathode terminals (through capacitor C_{block} Fig. 9-10) that the output voltage is taken.

Now assume that a negative video pulse appears at the grid. This pulse causes an abrupt decrease of anode current and the corresponding abrupt drop of voltage across resistor R_L . Simultaneously, anode voltage increases sharply by the same value. As a consequence, the blocking capacitor C_{block} starts

to charge, and the charging current flowing (see Fig. 9-10a) through resistor R_g from the top downwards sets up at its upper end a positive potential which acts on the grid of the succeeding valve.

When the pulse terminates, anode current increases abruptly, and anode voltage drops to its initial value. As this takes place, capacitor C_{block} ceases to charge through resis-

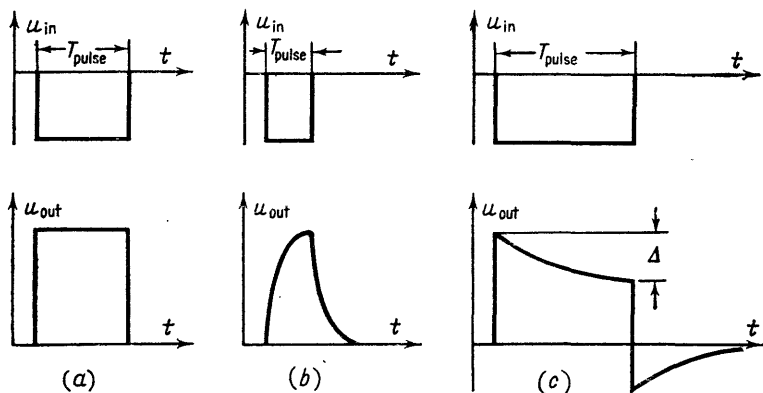


Fig. 9-23. Distortions in video pulse shape during amplification
(a) amplification without distortions; (b) distortion of short pulse; (c) distortion of long pulse

tor R_g and the current through resistor R_g also ceases so that the potential at its upper end becomes zero. Such an "ideal" (undistorted) process of video-pulse amplification is illustrated in Fig. 9-23a.

The behaviour described would be in the case if the shunt capacitance $C_{in\ 2}$ was zero and that of blocking capacitor C_{block} was equal to infinity. In other words, the process would be distortion-free if the frequency response did not decay at both the upper and lower frequencies. In an actual amplifier, however, pulse shape is distorted and the resultant distortions may be observed physically.

One of the reasons of pulse distortions is in that anode voltage cannot increase instantaneously but *rises gradually* as capacitance $C_{in\ 2}$ connected in parallel with resistor R_1 is being charged. The greater the product $R_1 C_{in\ 2}$, the slower is the charging process and the less the slope of the leading

edge of the output pulse (i. e., the greater its rise time) (Fig. 9-23b). Similarly, anode voltage cannot drop instantaneously since it is supported by the charge of capacitor $C_{in\ 2}$ which discharges gradually through resistor R_l and, therefore, the trailing edge of the output pulse is stretched (Fig. 9-22b). The shorter the pulse being amplified, the greater the danger of this kind of distortions (slow rising and falling of pulse edges).

The processes of capacitor charging and discharging through a resistor were described in Section 2-4. Specifically, we may use the voltage rise law

$$u_g = V_0 \left(1 - e^{-\frac{t}{\tau_u}} \right)$$

to describe mathematically the leading edge of the distorted pulse. Here, the rise-time constant $\tau_u \approx R_l C_{in\ 2}$ determines the passage of upper frequencies. For a pentode τ_u is dictated by the choice of R_l since $C_{in\ 2}$ cannot be less than the input capacitance of the valve. If it is required that the voltage should reach the 0.9 level of the stationary value V_0 in the time $0.1 T_{pulse}$, then the time constant should be taken as

$$\tau_u \approx \frac{T_{pulse}}{20} \quad (9-28)$$

where T_{pulse} is the pulse duration.

Assume, for instance, that the pulse duration $T_{pulse} = 1 \mu s$ and the capacitance $C_{in\ 2} = 50$ pF. Then from formula (9-28) we find

$$R_l \approx \frac{T_{pulse}}{20 C_{in\ 2}} = \frac{10^{-6}}{20 \times 50 \times 10^{-12}} = 10^3 \text{ ohms}$$

This value is *much less* than the resistance of load resistors used in audio-frequency amplifier stages and, consequently, the video-stage gain is not great. With $S = 5$ mS, the voltage gain at middle frequencies $K \approx S R_l = 5 \times 10^{-3} \times 10^3 = 5$.

Let us now turn to formula (9-18) and express the upper frequency in terms of $R_l \approx 1/G_T$ and $G_{in\ 2}$:

$$\omega_u \approx \frac{1}{R_l C_{in\ 2}}$$

In our example the upper frequency

$$\omega_u \approx \frac{1}{R_l C_{in\ 2}} = \frac{10^{12}}{10^3 \times 50} = 20 \times 10^6$$

This corresponds to frequency $f_u = 20/2\pi \approx 3.2$ MHz! Thus, we see that for the stage to amplify satisfactorily short pulses its *passband should be extended towards upper frequencies*, which is achieved, however, at the expense of a low voltage gain.

What determines then the lower limit frequency of the amplifier passband? To answer this question, we consider the amplification of a long pulse which is so long that its rise and decay times may be neglected (Fig. 9-23c).

In this case the pulse may be distorted due to a *decrease in the current* flowing through resistor R_g and in the *voltage across this resistor as capacitor C_{block} is being charged*. As a consequence, the pulse top becomes inclined (Fig. 9-23c). Furthermore, when the pulse terminates, capacitor C_{block} which has been charged during the pulse action is now discharged through resistor R_g . The discharge current flowing through resistor R_g from the bottom upwards produces a voltage undershoot across the resistor (Fig. 9-23c). This kind of distortion may be reduced by increasing the product $R_g C_{block}$ which is complicated because of cost, size, weight and insulation resistance of capacitor C_{block} as well as the necessity of ensuring charge leakage from the grid of the succeeding valve through resistor R_g .

The law of pulse top decay is also known from Section 2-4

$$u_{out} = V_0 e^{-\frac{t}{\tau_l}}$$

where $\tau_l = C_{block} R_g$ is the decay-time constant which determines the passage of lower frequencies. If we allow the voltage to fall by the value $\Delta = 0.1 V_0$ for the pulse duration period T_{pulse} , then we should take

$$\tau_l \approx 10 T_{pulse} \quad (9-29)$$

We know that R_g must not exceed $1 \text{ M}\Omega$, hence the capacitance $C_{block} \geq 10 T_{pulse} / R_g$. For a comparatively long video pulse $T_{pulse} = 1 \text{ ms}$ (i. e., 0.001 s) with $R_g = 10^6 \Omega$, the blocking capacitor should have the capacitance

$$C_{block} \geq \frac{10 T_{pulse}}{R_g} = \frac{10 \times 10^{-3} \times 10^6}{10^6} = 0.01 \mu\text{F}$$

cies. This method is called the *high-frequency compensation* of video amplifiers.

In modern transistorized television sets the video signals controlling the intensity of the kinescope beam are generally amplified by two wideband stages known as matching and output stages. These stages employ a common-collector and common-emitter circuit, respectively. It should be noted, however, that a wide passband is more difficult to be obtained in transistor stages than in the valve ones. This may be explained by a finite mobility of charge carriers in semiconductor crystals. Consequently, the amplifier passband is determined primarily by the intrinsic properties of the transistor rather than by the circuit parameters. The cutoff frequency (f_{cutoff}) which we discussed in Section 9-2 is the main of the transistor intrinsic parameters determining the operation of the common-emitter circuit.

High-frequency transistors are generally used in video stages. For instance, a transistor having a cutoff frequency of about 300 MHz is required for a video amplifier with the upper limit frequency $f_u=4$ MHz. Furthermore, the output stage of a video amplifier should provide a pulse signal with a peak amplitude of approximately some dozens of volts for successful electron beam modulation. Circuit correction (such as in the circuit shown in Fig. 9-24) is used, as a rule, for extending the amplifier frequency response.

Within the scope of this book this is practically all that can be said concerning primary signal amplifiers. Low-power amplifiers will be discussed in Chapter XII as applied to radio receivers; and power amplifiers, in Chapters X and XI.

Chapter X. WAVE GENERATION

10-1. OPERATING PRINCIPLES OF VALVE OSCILLATORS

For radio transmission a source of high-frequency continuous oscillations is required. It is quite natural to try to obtain these oscillations from special circuits in which they arise as a result of disturbance in electrical equilibrium.

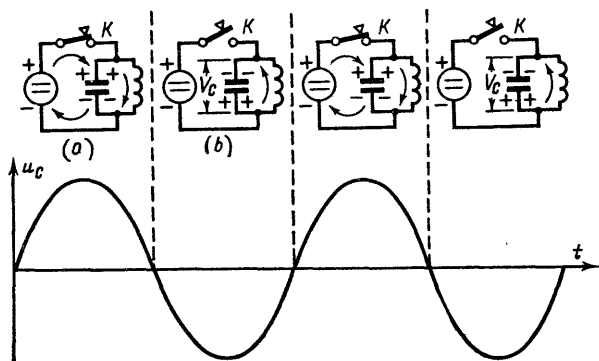


Fig. 10-1. Replenishment of circuit energy from DC source
(a) DC source replenishes circuit energy; (b) connection of DC source reduces circuit energy

An oscillatory circuit is a system capable of producing a sine-wave.

Let us connect a DC source to the oscillatory circuit at some instant t_0 for a very short time period by using switch K (Fig. 10-1). When instantaneously closing the circuit the current flowing through the coil (inductor) has no time to rise appreciably since the inductance tends to prevent any

changes in current. In other words, the right-hand branch of the oscillatory circuit will be essentially open. Simultaneously, the capacitor has time enough to be charged, i. e., to store some energy. When the DC source is switched off, the capacitor will discharge through the coil, and damped oscillations will occur in the circuit. To obtain sustained (continuous) oscillations, it is necessary to replenish periodically losses in the circuit. This is accomplished by connecting the DC voltage source to the oscillatory circuit at that part of the oscillation cycle when the capacitor plate connected to the negative terminal of the voltage source collects an excess of electrons. The source will then replenish the charge on the capacitor plates, i. e., restore the stock of energy in the capacitor (Fig. 10-1a). If the switch is closed at that half-cycle when positive charges are collected on the given plate, the source will neutralize them and discharge the capacitor. In the first case the electrons accelerated by the field of the DC voltage source will move to the circuit undergoing the counteraction of like charges collected on the plates of its capacitor. In the second case they will move in the circuit's accelerating field.

On the basis of this, we may formulate a common physical principle of energy exchange between the oscillating system and electron flow, which underlies the operation of most generators. *Energy is replenished in the oscillating system when electrons are supplied to it in a retarding field. In the accelerating phase the charges moving in the field of the oscillating system take away energy from the field.* This general rule characterizes the condition and direction of energy exchange between the electron flow and the field (see Chapter VIII).

It is clear from the above that energy in the oscillatory circuit can be replenished by connecting and disconnecting the DC source for alternating half-cycle periods (Fig. 10-1a and b). As this takes place, the energy of the DC source will be converted into energy of high-frequency oscillations in the circuit; this is known as *generation of high-frequency oscillations*.

Apart from the DC source, a switch is required to accomplish such commutation. Since we deal with high-frequency oscillations, one cannot possibly use mechanical switches. A three-electrode or multielectrode valve may serve as a re-

liable, high-speed, high-efficiency switch for many RF oscillators. Small changes in the voltage across the control grid make the valve conductive if it is in the cutoff condition. Response time of such a switch is determined by the time required for electrons to transit from the cathode to anode. If this time is much less than the oscillation period, the valve may be considered inertialess. Modern valves are practically inertialess up to frequencies measured by thousands of megacycles (an oscillation period of the order of 10^{-9} s).

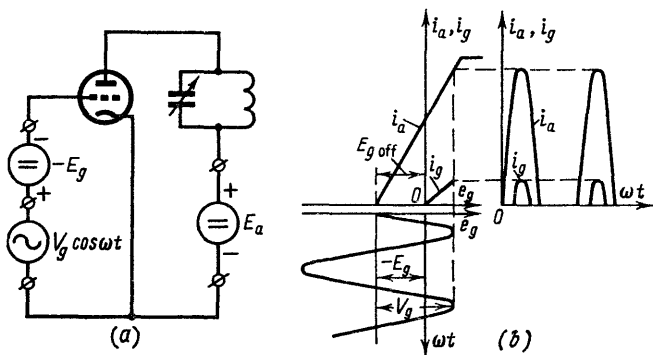


Fig. 10-2. Valve oscillator generating sustained oscillations
(a) schematic diagram; (b) currents and voltages in anode and grid circuits

To cut off the valve during the undesired half-cycle, a negative voltage, $-E_g$, called *bias voltage*, sufficient to stop anode current is applied to the valve control grid. The valve may be made conductive at the desired half-cycle by applying an additional AC *exciting voltage* $u = V_g \cos \omega t$ * (Fig. 10-2) to its control grid. Thus, the valve is cut off during negative half-cycles of exciting voltage and becomes conductive during positive half-cycles of this voltage. Therefore, under the action of resulting voltage across the control grid, which is

$$e_g = -E_g + V_g \cos \omega t \quad (10-1)$$

the current in the anode circuit will have the shape of periodic sine-wave pulses.

* In this chapter subscript max indicating peak amplitudes is omitted.

Such a device constitutes a *valve oscillator* producing high-frequency oscillations. To make sure that such an oscillator is operative, we should check two conditions: (1) whether current pulses will pass during the voltage phase favourable for energy replenishment in the circuit, and (2) whether oscillation power in the circuit will exceed the power drawn from the exciting source for valve control.

The first question may be answered by representing the anode current in the form of a sum of symmetrical-pulse harmonic components (see Section 2-5), i. e.,

$$i_a = I_{a0} + I_{a1} \cos \omega t + I_{a2} \cos 2\omega t + I_{a3} \cos 3\omega t + \dots \quad (10-2)$$

When the circuit is tuned to frequency ω , it constitutes practically a short circuit for the direct component and all higher harmonics. The voltage drop produced across it is only from the first harmonic, so we may write

$$u_t = i_{a1} R_{eq} = I_{a1} R_{eq} \cos \omega t \quad (10-3)$$

where R_{eq} is the circuit equivalent resistance.

According to Kirchoff's law for the closed network made up of the source, tuned circuit, and valve — anode voltage

$$u_a = E_a - u_t = E_a - I_{a1} R_{eq} \cos \omega t \quad (10-4)$$

In Fig. 10-3 voltages and currents in the oscillator are plotted as functions of time by using equations (10-1), (10-3)

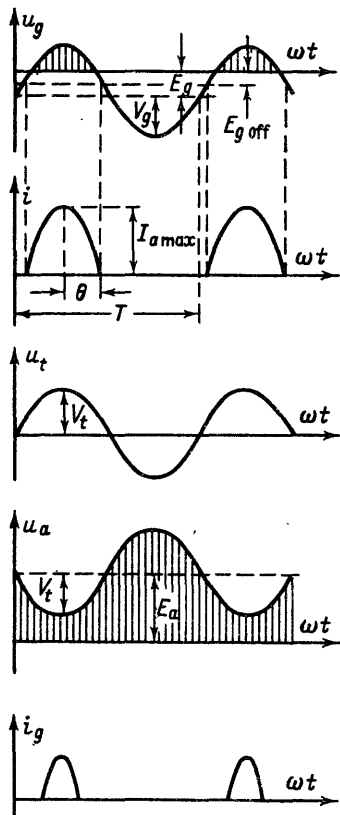


Fig. 10-3. Time diagrams of processes in valve oscillator

and (10-4). As is seen directly from these graphs, the electrons move in the valve under the influence of two fields, namely, the constant accelerating field of the anode voltage source and the tuned-circuit alternating field. The constant field accelerates the electrons by imparting kinetic energy to them. The alternating field, on the contrary, retards the electrons so that their velocity is decreased but energy in the oscillatory circuit is replenished at the expense of the kinetic energy lost by these electrons. In this way the energy of the DC source is converted into that of high-frequency oscillations.

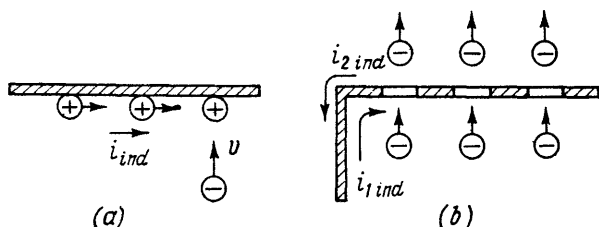


Fig. 10-4. Currents induced by moving charges
(a) current i_{ind} induced in conducting plane by approaching charge; (b) currents $i_1 ind$ and $i_2 ind$ induced in grid by approaching and departing charges

Let us now determine the power required to control electron flow in the valve. Assume first that the grid intercepts no electrons moving to the anode. This does not mean, however, that no current flows in the grid circuit.

If an electric charge (Fig. 10-4) appears near a conducting surface (in this case, the grid), free electrons in the conductor either collect or flow away from the region close to the charge depending upon its sign. This phenomenon is known as *electrostatic induction*, and the charges produced on the conducting surface are referred to as *induced charges*. Obviously, the amount of the induced charge depends upon the attracting force of the primary charge, i. e., upon the distance between the latter and the conducting surface. Therefore, if the primary charge moves to the surface, the induced charge will rise due to electrons which will move from other parts of the conductor, thus producing a unidirectional *induced current*. When the primary charge moves away from the surface, the induced charge will decrease in magnitude and elect-

rons will return to their initial position, thus producing an induced current of the opposite direction.

If the valve grid does not intercept electrons, as many electrons approach it as many move away from it to the anode, so that the resultant current I_{ind} induced in its circuit is zero. Therefore, the power of the exciting source which controls the flow of electrons in the valve will be in this case as follows:

$$P_{ex} = \frac{1}{2} V_g I_g = \frac{1}{2} V_g I_{ind} = 0$$

Consequently, if no electrons are intercepted by the valve grid, the exciting source will not spend power for electron-flow control (for a change in electron velocity).

This is a very important conclusion which explains the amplifying properties of most oscillators. It is precisely these properties that we talk about here since the exciting source spends generally some high-frequency power P_{ex} for electron-flow control, thus providing some power P of the same frequency in the anode circuit. An oscillator is said to exhibit amplifying properties if the first power is less than the second one, i. e., the oscillator should possess a *power gain* defined as

$$K_P = \frac{P}{P_{ex}} \quad (10-5)$$

and exceeding unity. In the ideal case discussed above the power gain is infinity.

What will be the power gain for an actual oscillator? Can we obtain practically such a gain? Strange as it may seem at first sight this is possible. Electrons will not settle down on the grid if the voltage applied to it is negative. Anode current, however, may exist under these conditions as it can be seen from valve static characteristics. Such operating conditions can be provided by the proper choice of bias and exciting voltages and are known as the *class A mode* (Fig. 10-5).

As seen from Fig. 10-5, however, this mode is characterized by low current efficiency and low power output. Therefore, the class A mode is used seldom in devices where it is important to obtain maximum output power, and the oscillators operate with control grid currents though their amplitude does not usually exceed 10 to 15 per cent of the anode current.

That is why the actual power gain of oscillators lies within dozens to hundreds of units.

As illustrated in Fig. 10-2, grid current is also a periodic pulse sequence, i. e., it consists of a direct component and harmonics, hence we may write

$$i_g = I_{g0} + I_{g1} \cos \omega t + I_{g2} \cos 2\omega t + I_{g3} \cos 3\omega t + \dots \quad (10-6)$$

To give the full idea of the valve oscillator operation, we should explain how the exciting voltage is produced.

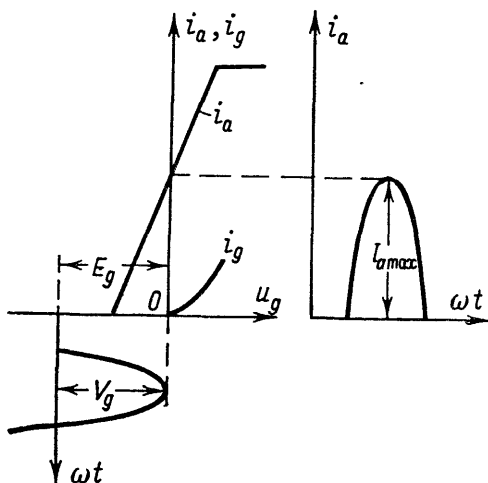


Fig. 10-5. Valve operating without grid currents (class A mode)

Often, another high-frequency oscillator may serve as the source of exciting voltage. In this case the oscillator being excited is called a *separately excited oscillator* (Fig. 10-6a). It is possible to eliminate the external exciting source by drawing a part of voltage from the oscillator anode circuit and applying it to the valve grid through a feedback path (Fig. 10-6b). Switching on or small disturbance of electrical equilibrium in such an oscillator results in oscillations, provided certain conditions are met. Therefore, such oscillators are generally called *self-excited oscillators* and are used as the sources of primary signals in radio equipment. The processes

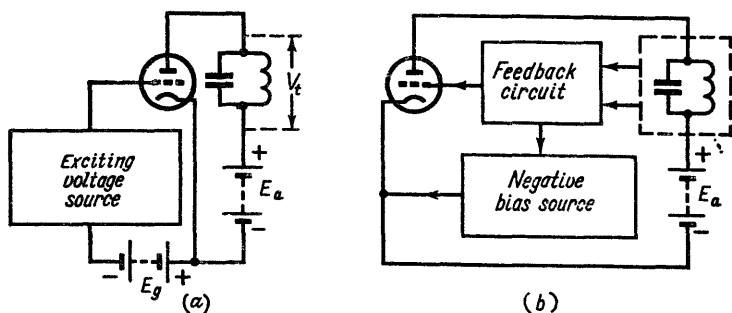


Fig. 10-6. Valve oscillators

(a) separately excited oscillator; (b) self-excited oscillator

occurring in them are more complex than in separately excited oscillators and, therefore, it is advisable first to study the latter in more detail.

10-2. SEPARATELY EXCITED OSCILLATORS (AMPLIFIERS)

From the above description of the general operating principles of valve oscillators it follows that the electron valve is used essentially as a *non-linear element* in them as distinct from the amplifiers discussed in Chapter IX. This means that during one part of the cycle the valve anode current changes proportionally to the voltage across the control grid and during the other part of the cycle anode current is absent although voltage across the grid still varies. As a consequence, the currents in the anode and grid circuits become non-sinusoidal, and only their individual components are used. This may seem irrational but in fact it is precisely the non-linearity of oscillators that, in principle, allows their efficiency to be obtained as close to unity as possible.

The mechanism of energy conversion in an oscillator discussed above allows one to conclude that this process becomes the more efficient the greater the force retarding the electrons moving in the valve. Therefore, in order to increase oscillator efficiency, electrons should move in the maximum retarding field and their number should be as great as possible. As

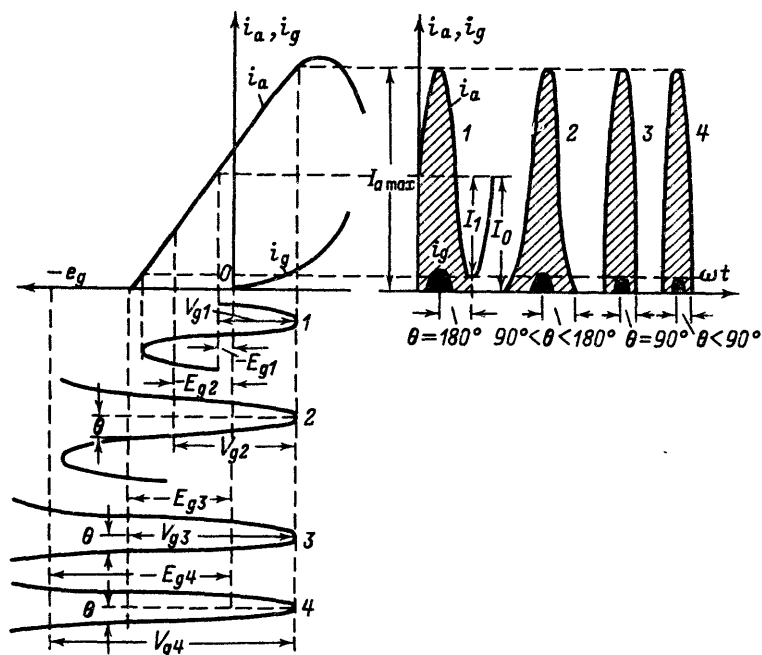


Fig. 10-7. Class A, B, and C oscillators (amplifiers)

seen from Fig. 10-3, to do this requires very narrow pulses of anode current.

The waveform of oscillator pulses is determined by their amplitude $I_{a \max}$ and the cutoff angle θ which is defined as half of that fraction of the cycle expressed in angular units during which anode current exists. In the case illustrated in Fig. 10-3, anode current exists during one half-cycle and, consequently, the cutoff angle is equal to 90° . Operation with the 90° cutoff angle is called *class B mode* (class B amplifier). When a valve operates with cutoff angles $\theta < 90^\circ$, its operation is referred to as *class C mode* (class C amplifier).

The above does not, however, imply that an oscillator cannot operate in a linear mode. In fact, the cutoff angle $\theta = 180^\circ$ can be obtained by selecting the valve bias voltage so that its anode current does not fall to zero (Fig. 10-7). In this case, the oscillator is said to operate in class A mode

(class A amplifier). The desired power effect is obtained in this mode even though the current flows continuously. This is explained by the fact that more electrons flow through it in the favourable (retarding) phase than in the unfavourable (accelerating) one. It is clear, however, that energy conversion efficiency in this case will be extremely low. Therefore, such a mode is used only when it is important to reproduce as accurate as possible the exciting voltage (input wave). This requirement should be met in most audio-frequency amplifiers.

Let us consider basic energy relations in a valve oscillator.

Proceeding from the conclusions of Section 2-4 that energy interaction of a voltage source takes place only between the currents of that frequency at which the voltage source operates, we may infer that the power drawn by the oscillator from the anode voltage source will be

$$P_0 = I_{a0} E_a \quad (10-7)$$

The useful power developed across the oscillatory circuit connected to the anode is defined as

$$P = \frac{1}{2} V_t I_{a1} = \frac{1}{2} \frac{V_t^2}{R_{eq}} = \frac{1}{2} I_{a1}^2 R_{eq} \quad (10-8)$$

Electrons not fully retarded by the oscillatory-circuit field strike the anode, spending the remainder of their kinetic energy for heating the anode. According to the energy conservation law, the power dissipated as heat in the anode is equal to

$$P_a = P_0 - P \quad (10-9)$$

One of the most important characteristics of any oscillator operating as an energy converter is the *electronic efficiency* defined as the ratio of the power generated to the power consumed, or

$$\eta = \frac{P}{P_0} = \frac{1}{2} \frac{I_{a1}}{I_{a0}} \frac{V_t}{E_a} \quad (10-10)$$

In compliance with expression (10-6) and conclusions of Section 2-4, the exciting source interacts only with the first harmonic of grid current having the same frequency as that of the source. Hence, the power of this source may be expres-

sed as

$$P_{ex} = \frac{1}{2} V_g I_{g1} \quad (10-11)$$

From equations (10-8) and (10-11), the power gain will be

$$K_p = \frac{P}{P_{ex}} = \frac{V_t}{V_g} \frac{I_{a1}}{I_{g1}} \quad (10-12)$$

From formulae (10-7) to (10-12) it follows that the energy relationships in a valve oscillator are determined by the harmonic components of valve currents which, in their turn,

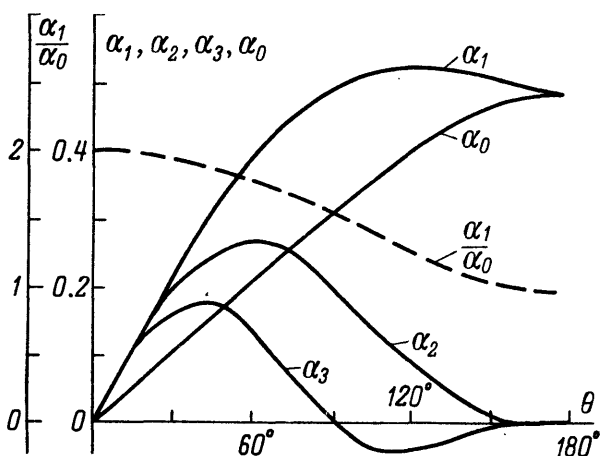


Fig. 10-8. Harmonic coefficients of sine-wave pulses as functions of cutoff angle

depend upon the waveform of pulses. Components of the sine-wave pulses used in valve oscillators are the functions of the cutoff angle.

In Fig. 10-8 the relative amplitude of three harmonics called pulse harmonic coefficients are shown as functions of the cutoff angle. These coefficients are

$$\alpha_0 = \frac{I_{a0}}{I_{a \max}} = f_1(\theta); \quad \alpha_1 = \frac{I_{a1}}{I_{a \max}} = f_2(\theta)$$

$$\alpha_2 = \frac{I_{a2}}{I_{a \max}} = f_3(\theta); \quad \alpha_3 = \frac{I_{a3}}{I_{a \max}} = f_4(\theta)$$

These graphs indicate that for each harmonic an optimum cutoff angle exists at which the fraction of the given harmonic in the pulse content will be a maximum. Optimum cutoff angles may be found from the simple equation

$$\theta_{opt} = \frac{120^\circ}{n} \quad (10-13)$$

The maximum values of harmonic coefficients decrease approximately in inverse proportion to the harmonic number. The anode-current direct component increases monotonically with the increasing cutoff angle.

By tuning the oscillatory circuit to the frequency of one of the higher harmonics, one may obtain oscillations whose frequency will be an exact integral multiple of the frequency of the exciting source. Such an oscillator is called a *frequency multiplier*.

By using adopted designations, expression (10-10) may be rewritten as follows

$$\eta = \frac{1}{2} \frac{\alpha_1}{\alpha_0} \frac{V_t}{E_a} \quad (10-14)$$

Figure 10-8 also illustrates the relationship between α_1/α_0 and the cutoff angle from which it is seen that this ratio equals unity for class A mode ($\theta=180^\circ$) and increases up to two as the cutoff decreases. If voltage across the circuit in class A mode is assumed equal to that of the exciting source (the extreme case for this mode A since with $V_t > E_a$ the voltage at the anode becomes negative at some part of the cycle, so that anode current ceases), then in compliance with formula (10-14) we obtain the maximum theoretical value for oscillator efficiency ($\eta_{\max}=0.5$). Decreasing the cutoff angle down to 90° at the same voltage across the circuit (class B mode) increases oscillator efficiency up to 78 per cent. A further decrease of cutoff (transfer to class C mode) leads to a monotonic increase of oscillator efficiency tending to unity as $\theta \rightarrow 0$.

From formula (10-14) it follows that the greater the ratio

$$\frac{V_t}{E_a} = \xi$$

the greater the oscillator efficiency. This ratio is called *anode efficiency*. The question arises what values of ξ are attainable

and advisable for practice. The graphs of the valve dynamic characteristics and anode current pulse waveforms answer this question.

As stated earlier, when the voltage at the grid changes according to the law $e_g = -E_g + V_g \cos \omega t$, it simultaneously leads, as distinct from the static mode, to voltage changes at the anode of the oscillator valve, which are described by

$$e_a = E_a - V_t \cos \omega t = E_a (1 - \xi \cos \omega t) \quad (10-15)$$

On the basis of this, one may calculate instantaneous grid and anode voltages at any instant of time for a given voltage across the oscillatory circuit (or for a given anode efficiency), and then find the corresponding magnitudes of anode and grid currents against the family of static characteristic curves. By performing such calculations for various instants of time, one may then draw a smooth curve connecting the separate points. The resultant line will represent an *oscillator dynamic characteristic curve* for the given voltage V_t across the circuit (Fig. 10-9a). Simultaneously, current pulse waveforms may be drawn (Fig. 10-9b).

The curves illustrated in Fig. 10-9 were plotted with the bias E_{gB} for a class B oscillator. In case of a short circuit in the anode network, $V_t = 0$ and $e_a = E_a$ at any instant of time (and at any voltage across the grid). Therefore, the dynamic characteristic curve is a vertical line 1 whose upper end terminates in the static characteristic corresponding to the maximum instantaneous voltage at the grid, which is

$$u_{g \max} = -E_g + V_g \quad (10-16)$$

It is easily comprehended that the dynamic characteristic curve 2 taken at a load impedance different from zero, i. e., with $V_t = V_{t2} > 0$, will be inclined since at the same grid voltages the instantaneous anode voltages, in compliance with formula (10-15), will be less than in case of a short circuit and, consequently, anode current will also be less.

The greater the voltage across the oscillatory circuit, the less the anode voltage and the less the slope of the dynamic characteristic curve. Owing to the small inclination of the horizontal portion of static characteristic curves, this does not lead to a substantial decrease in the pulse height until the upper end of the dynamic characteristic curve 3 intersects

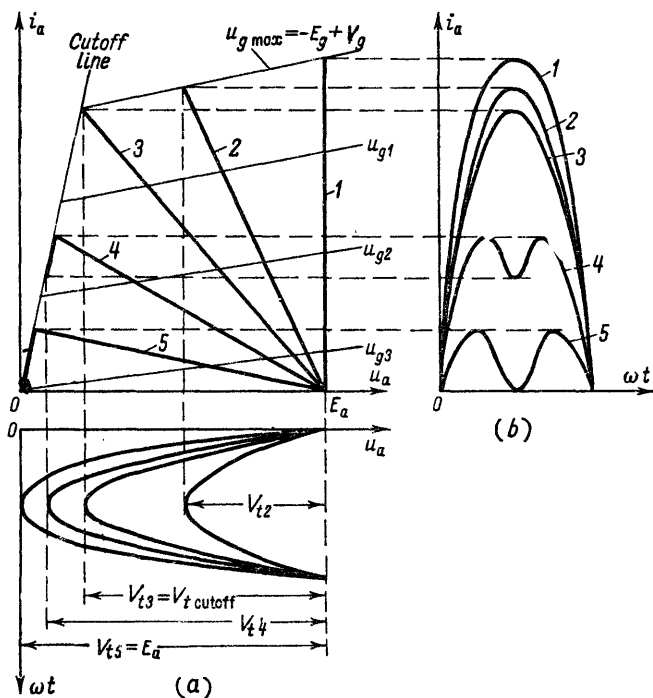


Fig. 10-9. Dynamic characteristic curves and anode current pulse waveforms for different voltages across oscillatory circuit

the drooping portion of the anode characteristic curve. The operating conditions corresponding to this case is known as *cutoff mode*.

A further increase in voltage across the oscillatory circuit results in that the upper end of the dynamic characteristic curve 4 descends along the drooping portion of the anode characteristic curve. As a consequence, the pulse height decreases and a valley appears at the pulse top. The last phenomenon may be explained by the sharp increase in grid current at the expense of anode current with small anode voltages. Such a mode is generally called an *overload* operation in contrast to the operating mode with low voltages across the circuit ($V_t < V_{t\text{ cutoff}}$), small grid currents and sharp

pulses of anode current, which is known as underload operation.

The anode-current pulse waveforms obtained with the aid of the graphical method make it possible to find the relationship between the harmonic components and voltages across the oscillatory circuit. All the components of anode current

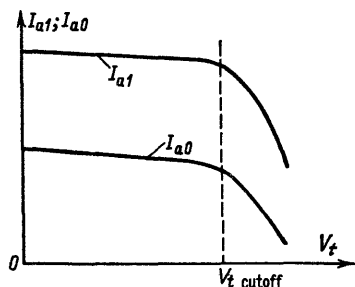


Fig. 10-10. First harmonic and direct component of anode current as functions of voltage across oscillatory circuit

are proportional to the pulse amplitude, therefore an increase in circuit voltage from zero up to $V_t \text{ cutoff}$ usually equal to (0.8 to 0.9) times E_a causes a small decrease in the direct component and the first harmonic of anode current (Fig. 10-10). A further rise in circuit voltage leads to their abrupt decrease in consequence of the reduction of anode-current pulse amplitudes and additionally because of the diminution of the content of these compo-

nents in pulses having a valley at their top.

The graph in Fig. 10-10 illustrates the relationship between the first harmonic of anode current and the voltage across the oscillatory circuit, i. e., $I_{a1} = f(V_t)$. There is, however, an inverse relationship, namely, oscillatory circuit voltage is proportional to the first harmonic of anode current, i. e.,

$$V_t = I_{a1} R_{eq} \quad (10-17)$$

Consequently, for a steady-state condition, current and voltage in the oscillator satisfy simultaneously these two relationships. They may be found as the points of intersection of the curves shown in Fig. 10-10 and the curves described by equation (10-17) as illustrated in Fig. 10-11. In the adopted coordinate axes the second relationship

$$I_{a1} = \frac{V_t}{R_{eq}} \quad (10-18)$$

is represented by straight lines whose angle of inclination decreases with the increasing value of R_{eq} .

By assigning different values to R_{eq} , we may obtain relationships between the current and voltage for different load resistances. These relationships are known as oscillator load characteristics. As seen from Fig. 10-11, an increase of load resistance from zero up to the value of $R_{eq} = R_{eq \text{ cutoff}}$ at which the oscillator goes into cutoff mode results in a nearly proportional increase of voltage across the oscil-

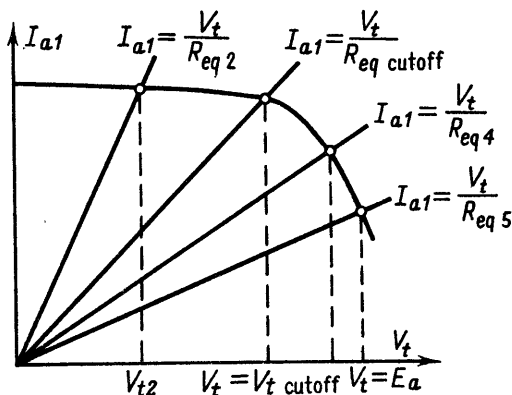


Fig. 10-11. Determining the voltage and current at given load resistance

latory circuit and in a slight decrease of anode current. A further rise of load resistance causes the oscillator to transfer to overload mode characterized by an abrupt fall of anode current with a slight increase in circuit voltage (Fig. 10-12). The output power, being the product of anode current and circuit voltage, rises and decreases during underload and overload modes, respectively—in the former, due to a voltage rise and, in the latter, because of an abrupt decrease in anode current. Consequently, we may conclude that the power being generated attains its maximum in the cutoff mode. As load resistance rises, the power being consumed decreases monotonically in compliance with the changes of the anode-current direct component. It is of interest that this is inherent only in oscillators operating in a non-linear mode. In class A operation the anode-current direct component does not depend upon the load resistance.

On the basis of the relationships obtained, we may find the oscillator efficiency change law. Oscillator efficiency increases up to the cutoff mode since the output power increases while the power being consumed decreases. In the overload region the oscillator efficiency remains almost constant owing to the fact that the power being generated and that being consumed decrease simultaneously. As seen

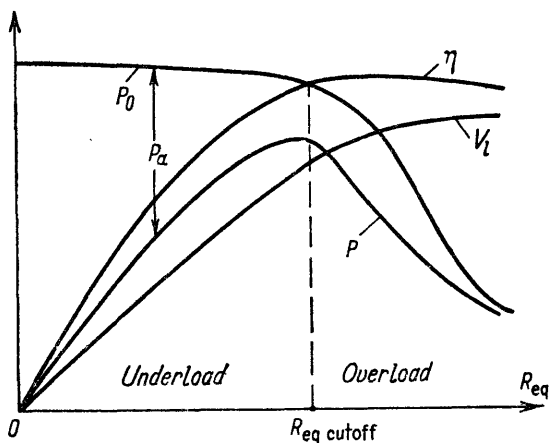


Fig. 10-12. Load characteristic curves of valve oscillator

from Fig. 10-12, the anode dissipation equal to the difference between the power being consumed and the output power decreases monotonically with the increasing load resistance.

The above discussion shows that the maximum power with a high efficiency may be obtained when the oscillator is used in the cutoff mode corresponding to $V_t = (0.8 \text{ to } 0.9) E_a$.

We now consider how the operation of a grid circuit changes with changing load resistance. As load resistance is increased, voltage across the load rises, whereas the minimum anode voltage corresponding to the maximum grid voltage drops (Fig. 10-13). Therefore, valve electron flow is redistributed in favour of the grid. Grid current increases comparatively slightly and very abruptly in the underload and

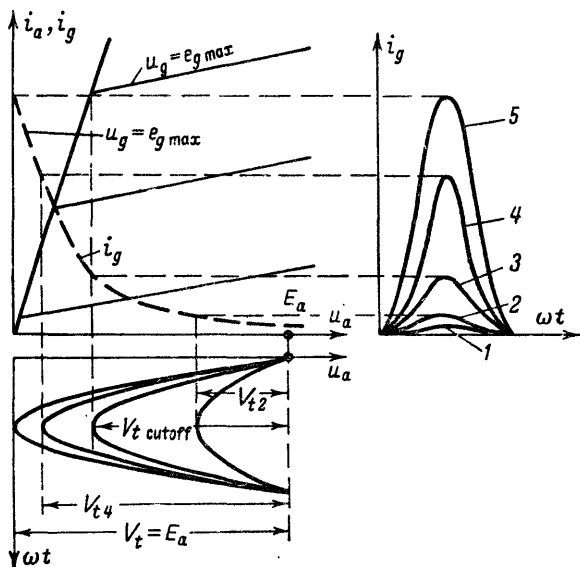


Fig. 10-13. Grid-current pulse waveform as a function of voltage across oscillatory circuit

overload modes, respectively. Power drawn from the exciting source increases simultaneously.

From the schematic diagram of an oscillator grid circuit shown in Fig. 10-14a, it can be easily seen that the direct component of grid current I_{g0} recharges the bias source E_g , with the corresponding power being equal to

$$P_{g0} = I_{g0} E_g \quad (10-19)$$

Since grid current is produced under the action of the exciting source, the bias source is recharged at the expense of power P_{ex} drawn from the exciter. The remainder of this power is dissipated in the form of heat in the grid, thus, we may write that

$$P_g = P_{ex} - P_{g0} \quad (10-20)$$

From the above discussion it follows that anode and grid overheating must be avoided when using a low load resistance (underload mode) and a high load resistance (overload

mode), respectively. In triodes overheating is a serious hazard to the control grid, but in tetrodes and pentodes overheating constitutes danger to the screen grid as well since an appreciable positive potential is applied to it during operation.

All the above refers to oscillators whose grid bias is produced by a separate supply source and does not change with changing operating conditions. Generally, this kind of bias is used rather seldom. In fact, why should a supply source

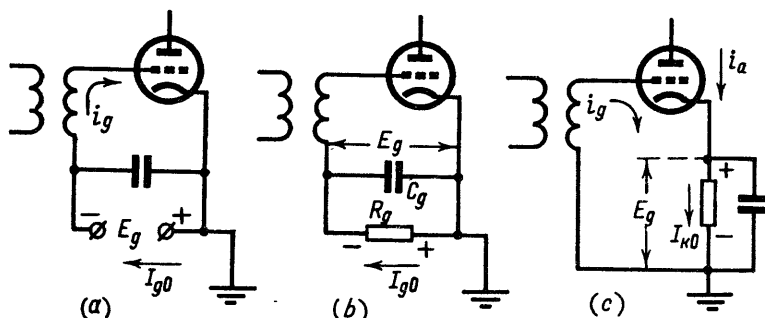


Fig. 10-14. Methods used for providing negative bias voltages across control grid of valve oscillator

(a) connecting the grid bias source; (b) grid bias; (c) cathode bias

be employed if its power is not used? Therefore, it is most frequently replaced by a network consisting of resistor R_g and capacitor C_g connected in parallel (Fig. 10-14b). While passing through the resistor, the grid-current direct component produces a voltage drop across it which is equal to

$$E_g = -I_{g0}R_g \quad (10-21)$$

The resulting voltage drop serves as the grid voltage negative with respect to the cathode.

As grid current alternating components pass through capacitor C_g having a rather high capacitance, they do not produce an appreciable voltage drop across it, and one may consider that voltage across the network remains essentially unchanged. Such an elementary network is known as an *automatic bias network* or a *self-bias network*. It should be noted that this network possesses a number of interesting

features associated with oscillator operation. Grid current changes with oscillator operating conditions, so that grid bias also varies. At the same time a separate bias source offering a low internal resistance maintains a constant bias irrespective of the oscillator operating conditions.

Application of an automatic bias network may be useful in a number of cases. For instance, when the oscillator operating conditions approach those of the overload mode because of changes in supply voltages, grid current will increase and bias will rise so that the overload will somewhat reduce. In other words, the bias network stabilizes operating conditions (oscillator loading).

However, the automatic bias network in some instances may cause unwanted effects. For instance, if the anode circuit is mistuned, voltage across it will drop, and the anode residual voltage, $e_{a \min} = E_a - V_t$, will rise. As a result, anode current will increase and grid current decrease. This, in turn, leads to a bias decrease and an additional rise of anode current causing anode overheating.

A *cathode bias network* is frequently used to avoid the above disadvantages (Fig. 10-14c). It is similar by its operating principle to the grid bias network but bias is produced here by the cathode-current direct component. As cathode current represents the total current of all the electrodes, its magnitude does not depend upon changes in oscillator loading.

Combined bias circuits in which bias is produced simultaneously by grid and cathode currents are often used. Such combined circuits assure high stability of oscillator operating conditions.

Oscillator operation is also affected by the method used for supplying power to other valve electrodes. In the schematic diagrams discussed above the anode was supplied through the oscillatory circuit. This is dangerous if an operator should adjust the circuit tuning during operation. Such an anode supply method is known as a *series feed*. To eliminate danger to the operator's life, a *parallel feed* is generally used to supply the anode in oscillators operating with anode voltages exceeding 250 V (Fig. 10-15). The anode is separated here from the oscillatory circuit by blocking capacitor C_{block} having a capacitance high enough to present a short circuit for high-frequency currents. In its turn, the anode

voltage source is connected through blocking choke L_{block} which prevents high-frequency currents being shorted through the supply source. The reactance of the choke for the first harmonic of anode current should be approximately ten times as great as the value of the oscillatory circuit equivalent resistance. Since the circuit equivalent resistance amounts to thousands of ohms, the choke should have a large number of turns.

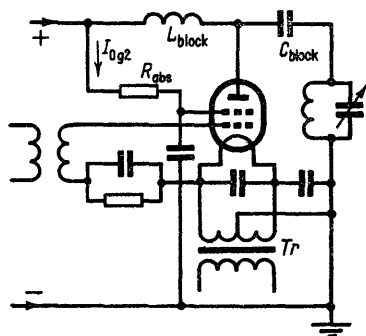


Fig. 10-15. Power supplies for an oscillator valve

Here, the suppressor grid is DC connected to the cathode or a small positive voltage is applied to it (usually from a potentiometer inserted into the common supply circuit).

The valve screen grid is supplied with an appreciable positive voltage comparable with the anode voltage. Therefore, this grid is generally fed from the anode supply source through the absorbing resistor R_{abs} (Fig. 10-15). The direct component of the screen grid current, I_{0g2} , produces the voltage $V_{abs} = I_{0g2}R_{abs}$ across the absorbing resistor, thus, the screen grid voltage will be

$$E_{g2} = E_a - I_{0g2}R_{abs} \quad (10-22)$$

When such a supply method is used, any variations in oscillator loading as in the case of the automatic bias network change the screen grid voltage in such a way that oscillator operating conditions are stabilized.

Filament (heater) voltage is most frequently taken from a step-down transformer, Tr , one of the ends of which or the secondary winding centre-tap is grounded. The last circumstance should always be borne in mind since the valve cathode in some circuits is under high-frequency voltage relative to the ground. High-frequency chokes should be inserted in this case into the heater circuit. In high-power oscillators employing valves with directly heated cathodes high-frequency currents may produce an appreciable potential

difference across the filament. To eliminate the hazard of electric break-down the high-frequency potentials of filament ends are generally made equal by connecting the filament leads to each other through a capacitor.

10-3. SELF-EXCITED OSCILLATORS

As distinct from the separately excited oscillators discussed above, self-excited oscillators are provided with a feedback path. Fig. 10-16 illustrates a schematic diagram of a self-oscillator in which feedback between the anode and grid of the valve is accomplished through a transformer.

The presence of feedback makes it possible to excite sustained oscillations in the tuned circuit without an external source. This may lead to misunderstanding since if oscillations existed in the circuit at the initial moment, exciting voltage would arise at the grid due to the presence of feedback and oscillations in this closed system would continue indefinitely owing to the amplifying properties of the oscillator. But no oscillations existed in the system at the instant when the oscillator was switched on.

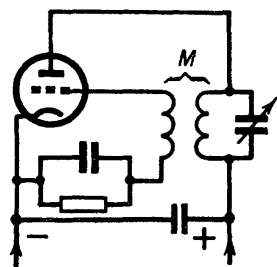


Fig. 10-16. Self-oscillator with transformer feedback

Harmonic analysis clears this misunderstanding. When the oscillator is switched on, momentary current pulses appear in all the oscillator circuits, which charge the circuit capacitances. As is known from Section 2-5, a single pulse has a continuous frequency spectrum and one of its frequency components will coincide with the natural frequency of the oscillatory system. This oscillation will excite the tuned circuit and voltage of the given frequency will be fed to the valve grid through the feedback path. This voltage will cause the anode current to change with the same frequency, and its AC component while passing through the tuned circuit will amplify the oscillations initiated in it. The amplitude of these oscillations will rise until the power delivered to the tuned circuit is balanced with the increasing

losses, so that steady-state oscillations with a constant amplitude are set up in the circuit. This process is called *oscillator self-excitation*. Since the oscillator in this case presents a self-contained system (free from external effects), it is generally referred to as a *self-oscillator* and the oscillations generated by it are called *self-oscillations*.

It should be noted that the process of oscillator excitation outlined above is somewhat simplified. In fact, some conditions should be satisfied so that the oscillator can operate in this fashion.

First, to amplify primary voltage initiated in the tuned circuit, the alternating component of anode current should produce a voltage drop across it which is in phase with the primary voltage, i.e., the sum of all phase shifts accumulated when passing around the closed oscillatory system should be zero or equal to an integral number of 2π rad. This requirement is known as a *phase condition of self-excitation* or a *phase balance condition* for a self-oscillator. This condition determines the frequency of the oscillations being generated since under the given circuit parameters phase shifts depend upon frequency.

Second, primary oscillations will increase if the AC power delivered to the tuned circuit exceeds the losses in it. This condition is satisfied under certain relationships between the amplitudes of currents and voltages in the oscillator and, therefore, it is known as an amplitude condition of self-excitation. Let us now express this condition mathematically. In compliance with the design formulas derived in Section 8-4 the first harmonic of anode current may be written in the following form:

$$I_{a1} = S(V_g + DV_a) \quad (10-23)$$

where S is valve transconductance; V_g , grid voltage; V_a , anode voltage, and D , valve grid-through ($D=1/\mu$).

Grid voltage is proportional to voltage across the tuned circuit (output voltage), i.e.,

$$V_g = kV_t \quad (10-24)$$

where k is the *feedback ratio* indicating what portion of output voltage is fed back to the grid.

Alternating voltage across the valve anode is out of phase with that across the tuned circuit (see equations (10-4), (10-3) and Fig. 10-3)

$$V_t = -V_a \quad (10-25)$$

Substituting expressions (10-24) and (10-25) in formula (10-23), we obtain

$$I_{a1} = S V_t (k - D) \quad (10-26)$$

Voltage produced by this current across the tuned circuit is

$$V'_t = I_{a1} R_{eq} = S V_t (k - D) \quad (10-27)$$

where R_{eq} is the circuit equivalent resistance, and V'_t , voltage developed in the tuned circuit under the action of primary voltage V_t . If the power delivered to the tuned circuit exceeds the losses in it, then

$$V'_t > V_t \quad (10-28)$$

and oscillations will rise.

Substituting formula (10-27) into (10-28), we have

$$S R_{eq} (k - D) > 1 \quad (10-29)$$

or

$$k > \frac{1}{S R_{eq}} + D \quad (10-30)$$

Hence, it is clear that an oscillator can be self-excited only if the feedback ratio exceeds some critical (minimum) value which is the greater, the less the valve transconductance and circuit equivalent resistance and the greater the valve grid-through.

To obtain high efficiency, the oscillator should be operated either as class B or class C. For this purpose, an appreciable negative bias, $-E_g$, which cuts off the valve in the absence of exciting voltage must be applied to the grid. In this case small oscillations will not make the valve conductive (Fig. 10-17) and the oscillator will not get excited. The oscillator can be self-excited if oscillations of rather high amplitude are produced in the circuit when the oscillator

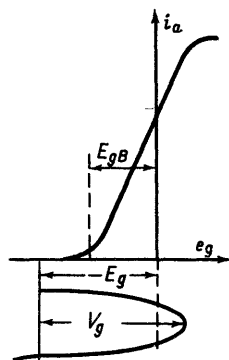


Fig. 10-17. Bias voltage at valve control grid in class C operation

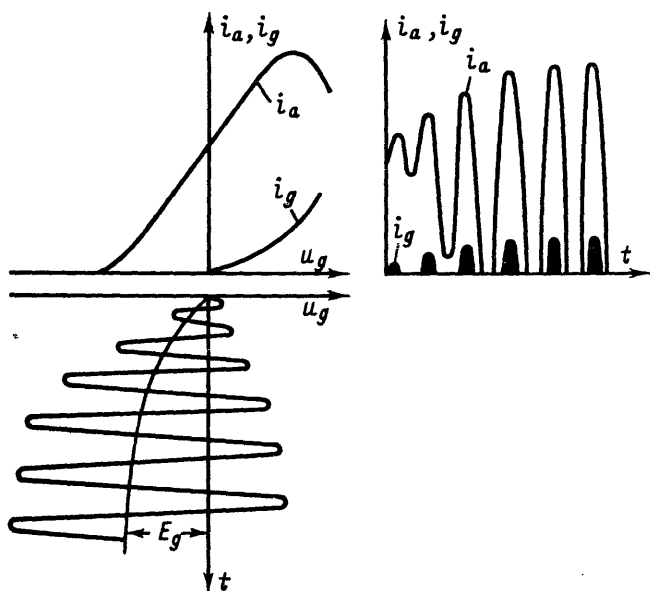


Fig. 10-18. Voltages and currents in process of self-excitation of self-oscillator provided with automatic grid bias.

is switched on (such a self-excitation mode is generally referred to as the "rigid" operation). This condition, however, is not usually satisfied. Therefore, the valve must be made conductive at the initial moment and operate in the maximum transconductance region as follows from equation (10-30) in order that even small electric disturbances could initiate oscillations in the oscillator. For this purpose, grid bias should be either zero or very small. This means, however, that the oscillator will have a very low efficiency.

The application of automatic grid bias helps to eliminate the arising contradiction between the requirements for high efficiency and easy or, as it is customary to say, "soft" self-excitation. At the instant the oscillator is switched on, no oscillations exist as yet and, thus grid voltage, grid currents and, consequently, bias are also absent, so that the valve is in the region of high transconductance (Fig. 10-18) and the oscillator is easily excited. Exciting voltage, grid current

and the bias voltage produced across the $R_g C_g$ network increase with the increasing amplitude of oscillations. A sufficiently large bias is produced during steady-state operation, and the oscillator is operated as class *B* or *C* characterizing by high efficiency. For these reasons, the automatic bias network is used practically in all self-oscillators.

We now consider in more detail why oscillations stop to increase and how the oscillator obtains its steady-state operation with constant amplitudes. While inspecting Fig. 10-18, one may note that an increase in bias during oscillator excitation generally results in that an ever increasing part of the cycle of grid alternating voltage u_g falls within those portions of the valve characteristic curve where transconductance is either heavily reduced or zero. Therefore, it is advisable to speak about some *average transconductance* which decreases as the cutoff angle increases. For instance, in class *B* operation the valve is cut off during one half-cycle and, therefore, average transconductance is twice as less as the static transconductance of the conductive valve. A decrease in average transconductance results in the reduction of the amplitude of the anode-current first harmonic and the power delivered to the tuned circuit.

On the other hand, an increase in grid voltage and grid current leads to a rise of the exciting power which is taken from the tuned circuit. As this takes place, the losses inserted into the tuned circuit increase, whereas circuit equivalent resistance drops. An increase in exciting voltage beyond a predetermined magnitude may stop the anode voltage rise because of transition to the overload mode in which the content of the first harmonic is decreased as a result of distortions in pulse waveform. This may also be considered as a decrease in average transconductance.

Oscillations cease rising when the power delivered to the circuit and that consumed in it are balanced. Power balance is characterized by the equality of the circuit primary voltage and the voltage produced under its action, i.e.,

$$V'_t = V_t \quad (10-31)$$

From this equation with allowance made for (10-27), we obtain

$$k = \frac{1}{S_{av} R_{eq}} + D \quad (10-32)$$

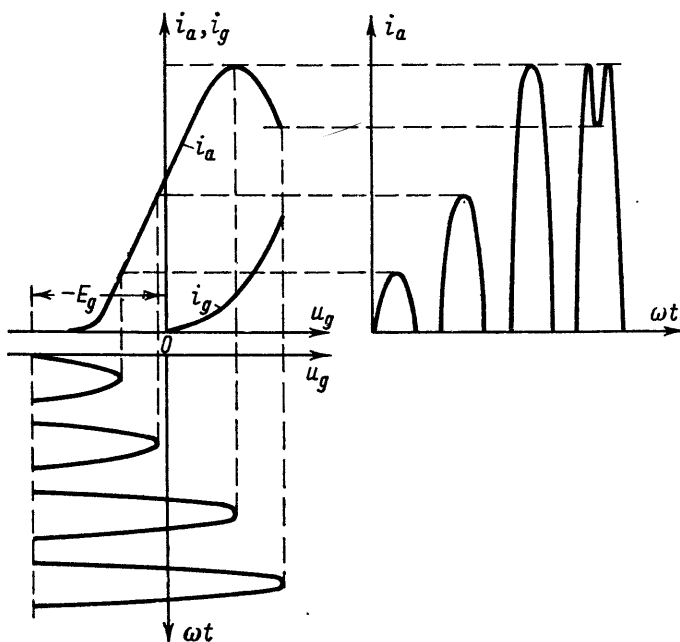


Fig. 10-19. Relationship between anode current and grid voltage for oscillator with constant bias

The above relationship expresses the condition for steady-state oscillations in the oscillatory system. As it characterizes energy balance in an oscillatory system, we shall call it as an *amplitude-balance equation*.

Let us now consider the selection of the feedback ratio and its effect on the oscillation amplitude. The answer to this question may be obtained by considering the so-called "oscillatory characteristics" of oscillators. These characteristics are based on the fact that a self-oscillator may always be represented as a separately excited oscillator whose input and output are coupled through a feedback path.

Oscillatory characteristics are the relationships between the first harmonic of anode current in an open-feedback oscillator and the exciting voltage taken for several values of grid bias E_g . When $E_g=0$, the valve is conductive, and

the first harmonic of anode current rises proportionally to an increase of alternating voltage applied to the grid until an increase in grid current stops this process. As a consequence, the first harmonic of anode current will even decrease because of a valley appearing on the pulse top (Fig. 10-19).

If the experiment is repeated with a small negative bias, $-E_{g1}$, applied to the grid, anode current will increase first slowly due to the non-linear lower portion of the valve characteristic curve, and then it will rise faster. This rise, however, will soon come to a stop and current will fall as the valve is transferred to overload mode. Owing to the presence of some bias, this will occur at a higher exciting voltage than in the first case. The same changes will be more pronounced as bias is increased further ($-E_{g2}$, $-E_{g3}$, etc.). As a result of these experiments, we may obtain the characteristic curves shown in Fig. 10-20.

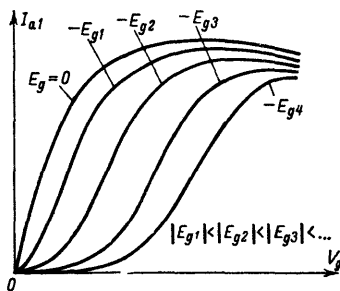


Fig. 10-20. Oscillatory characteristic curves

It will now be recalled that due to the presence of feedback in a self-oscillator there also exists an inverse relationship, namely, the greater the anode-current first harmonic, the greater the voltage across the tuned circuit and the exciting voltage which is proportional to the latter. Thus, we can write

$$V_g = kV_t = kI_{a1}R_{eq} \quad (10-33)$$

If we plot these relationships in the coordinates (I_{a1} , V_g) adopted for oscillatory characteristics curves, they will have the form of a family of straight lines running from the coordinates origin, with their angle of inclination being the greater, the greater the feedback ratio, i.e.,

$$I_{a1} = \frac{V_g}{kR_{eq}}$$

The point of intersection of the feedback line with the oscillatory characteristic curve (Fig. 10-21) determines the

amplitude of the anode-current first harmonic and the grid voltage satisfying both relationships. The values obtained are set up in the oscillator at a given feedback ratio and bias.

As seen from Fig. 10-21, when $k < k_{crit}$, feedback lines do not intersect oscillatory characteristic curves, thus indicating that the oscillator cannot be self-excited. When $k > k_{crit}$

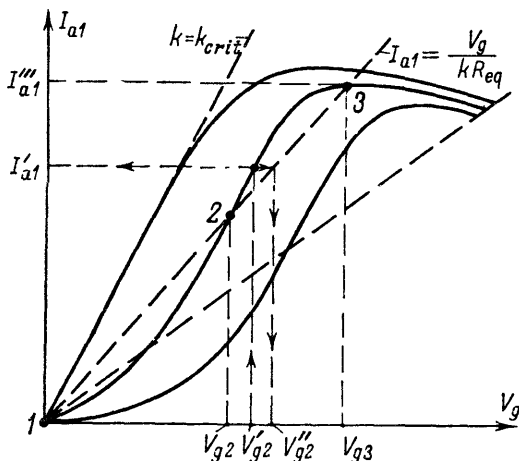


Fig. 10-21. Determining the steady-state conditions from oscillatory characteristic curves

self-oscillation becomes possible but in a number of cases there are three intersection points (1, 2, 3). The question is which of these three points determines the steady-state condition of the oscillator?

It can easily be shown that point 2 will be unstable. In fact, assume that grid voltage has increased from V_{g2} to V'_{g2} due to some accidental causes. As a consequence, anode current increases up to the value I'_{a1} . On the other hand, we see from the feedback line that grid voltage in this case will be $V_{g2} > V'_{g2}$, etc. Current and voltage will increase quickly up to the values I'''_{a1} and V_{g3} which will prove to be stable since an increase in current leads to a decrease in voltage and vice versa. Similarly, we may make sure that even a slight decrease in current or voltage at point 2 leads to a quick

transfer of the oscillator into a state corresponding to point 1, i.e., to oscillation disruption.

As seen from the same figure, an increase of feedback ratio from the value k_{crit} results first in a rise of anode current and grid voltage; this rise then ceases and even some fall of anode current occurs because the oscillator is transferred to overload mode. Consequently, to excite the oscillator, the feedback should exceed some critical value. Feedback, however, should not be chosen too strong as this leads to a decrease rather than to an increase in oscillator output power. Optimum feedback is found either by calculation or empirically.

Most valve self-oscillators are not provided with a special feedback element such as the coupling coil in the above example (see Fig. 10-16). The voltage applied to the valve grid is simply taken from one of the oscillatory system components by connecting the grid-cathode portion of the valve directly to this component. Consequently, in such oscillators the valve is connected to the oscillatory system at three points which are the anode, cathode and grid. Therefore, this type of oscillators may be referred to as a *three-point* configuration.

For convenience in studying the great variety of three-point oscillators let us familiarize ourselves with their general properties. Assume that the oscillatory system reactances (Fig. 10-22) connected between the valve anode and cathode, between the cathode and grid, and between the grid and anode are designated as X_1 , X_2 and X_3 , respectively. (Resistive components may be neglected as compared with reactive ones.) In general, these reactances may represent equivalent parameters of any complex network. Let us also consider that they include all stray parameters of the circuit such as lead inductances, interelectrode capacitances, etc.

An oscillator may generate waves only at the natural frequency of its oscillatory system. It will be recalled that in such a system the magnetic energy being stored is pe-

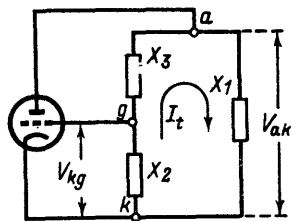


Fig. 10-22. Three-point self-oscillator

riodically transferred to electrical energy and vice versa. The amounts of these energies are equal. Therefore, a lossless oscillatory system does not consume energy. Consequently, for the total current I_t flowing around the oscillatory system the sum of reactances is zero, i.e.,

$$X_1 + X_2 + X_3 = 0 \quad (10-34)$$

As all the reactances depend upon frequency, equation (10-34) determines the frequencies of the oscillations which may be excited in the oscillator. Depending upon the kind of an oscillatory system, there may be one, two, three or, in the general case, an infinite number of such frequencies.

Voltage between the anode and cathode, V_{ak} , may be determined as the product of current I_t in the oscillatory system and reactance X_1 connected between these electrodes. Thus, we may write that

$$V_{ak} = I_t X_1 \quad (10-35)$$

Similarly, voltage between the cathode and grid will be

$$V_{kg} = I_t X_2 \quad (10-36)$$

Exciting voltage, on the contrary, is read from the grid to cathode and is defined as

$$V_{gk} = -V_{kg} = -I_t X_2 \quad (10-37)$$

From expressions (10-35) and (10-36) we obtain that

$$V_{gk} = -V_{ak} \frac{X_2}{X_1} \quad (10-38)$$

From the graphs shown in Fig. 10-3 it follows that in order to replenish energy losses in the tuned circuit, exciting voltage should be out of phase with the voltage between the anode and cathode. This condition will be satisfied if

$$\frac{X_2}{X_1} > 0 \quad (10-39)$$

From (10-38) it also follows that this ratio expresses the feedback ratio for the oscillator. In fact,

$$\frac{X_2}{X_1} = -\frac{V_{gk}}{V_{ak}} = k \quad (10-40)$$

The condition (10-39) is fulfilled if both reactances X_2 and X_1 are either positive or negative. Therefore, to satisfy condition (10-34), reactance X_3 should be of opposite sign to X_1 and X_2 .

Let us now consider possible self-oscillator circuits satisfying the above requirements. If $X_1 > 0$, $X_2 > 0$, and $X_3 < 0$, the self-oscillator equivalent circuit will take the form as illustrated in Fig. 10-23a. Such a circuit configuration may be called a *three-point self-oscillator with inductive feedback*.

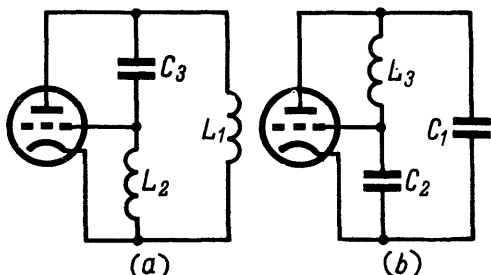


Fig. 10-23. Single-circuit three-point self-oscillators

(a) self-oscillator with inductive feedback; (b) self-oscillator with capacitive feedback

(Other names are a tapped-coil oscillator and Hartley oscillator.) If $X_1 < 0$, $X_2 < 0$ and $X_3 > 0$, the self-oscillator equivalent circuit will be as in Fig. 10-23b. This circuit configuration is called a *three-point self-oscillator with capacitive feedback*. (Other names are a tapped-capacitor oscillator and Colpitts oscillator.)

A knowledge of general self-excitation conditions and properties of a self-oscillator helps in understanding the circuits themselves and operating features of different practical circuits.

Single-circuit self-oscillators. Among the simplest self-oscillators circuits are configurations in which reactances X_1 , X_2 and X_3 form one oscillatory (tuned) circuit (tank circuit). Such self-oscillators may be called *single-circuit self-oscillators*.

Figure 10-24a illustrates a principle circuit of a single-circuit self-oscillator with inductive feedback. In this cir-

cuit inductors L_1 and L_2 are two portions of one inductance coil from one part of which voltage is fed back to the grid as in the case of an autotransformer. Therefore, this circuit may also be called an *autotransformer-feedback oscillator*.

One should pay attention to the fact that point k indicating the cathode connection should be located exactly between points a and g which are anode and grid connection points. Only in this case is proper phasing of the circuit

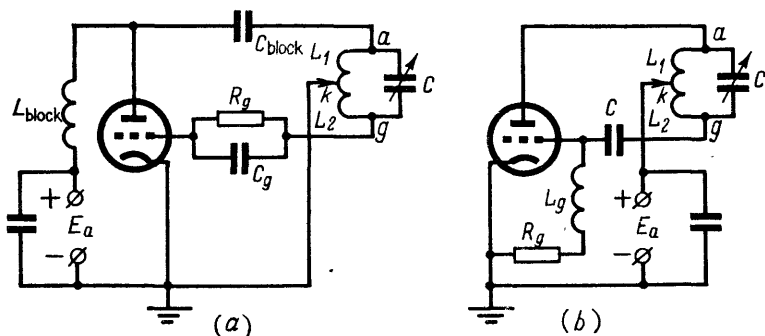


Fig. 10-24. Autotransformer-feedback self-oscillators (coil-tapped oscillators)

provided (grid voltage changes will be out of phase with anode voltage). The circuit will not be capable of self-excitation if the cathode and grid connection points are interchanged. In the corresponding equivalent circuit, in compliance with the designations adopted, capacitor C plays the role of reactance X_1 , and reactances X_2 and X_3 will be presented by the portions L_1 and L_2 of the inductance coil. Such a circuit is not capable of self-excitation since its feedback ratio is

$$k = -\frac{\omega L_2}{\frac{1}{\omega C}} < 0$$

As seen from Fig. 10-24a, the grid-current direct component path is closed through the portion L_2 of the inductance coil. Since the automatic bias network is connected in series, we should invariably use the parallel feed for anode supply, otherwise anode voltage will be applied to the grid.

This may be avoided, however, if an automatic-bias parallel network is used (Fig. 10-24*b*).

Feedback may be adjusted by moving point *g* along the inductance coil. Any changes in the position of point *k* result in simultaneous changes of the feedback ratio and tuned-circuit tapping factor, i.e., in the changes of equivalent load resistance. Oscillation frequency is most often varied by capacitor *C*.

In such circuits frequency can be varied continuously with overlap $k_f = f_{\max}/f_{\min}$ being of the order of 10. Such

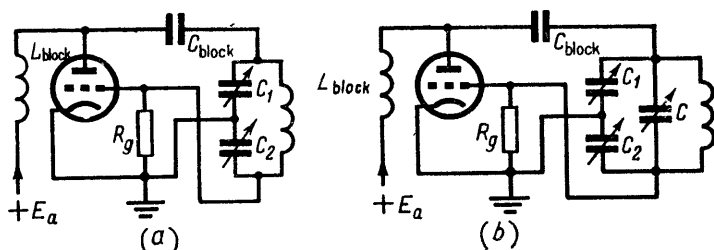


Fig. 10-25. Self-oscillators with capacitive feedback (capacitor-tapped oscillators)

oscillators have found wide application in the long- and medium-wave ranges. At frequencies above 10 MHz, coil inductance is reduced so much that only a few turns are required in the coils. As a consequence, feedback adjustment becomes extremely coarse and, therefore, it is advisable to use other types of self-oscillator circuits for the frequencies indicated.

Figure 10-25*a* illustrates a practical circuit of a three-point self-oscillator with capacitive feedback. Feedback voltage is taken from capacitor C_2 . Such an oscillator requires an automatic-bias parallel-network and the parallel feed for anode supply. As compared with the preceding circuit, the disadvantage of the capacitive-feedback circuit consists in the fact that feedback adjustments made by means of capacitor C_2 lead invariably to changes in oscillator frequency. To reduce this effect, capacitances C_1 and C_2 are chosen as low as possible and the circuit is tuned with the aid of an additional large-capacitance capacitor C (Fig. 10-25*b*). This oscillator also operates well at long and medium waves. At shorter waves circuit stray capacitances (such as

interelectrode capacitances, wiring capacitance) become comparable with circuit capacitances. This involves difficulties in feedback adjustment and narrows the operating range of oscillator frequencies.

The major advantages of single-circuit oscillators are the simplicity of design and tuning and a wide tuning range. Their principal disadvantage is that the only tuned circuit determining the operating frequency also serves as the oscillator load in which the power being generated is developed. Moreover, the same circuit is coupled to an external load. Since the power developed in the tuned circuit is converted into heat, circuit components change their sizes and electrical parameters, thus causing drift of oscillator frequency. An external load, in its turn, inserts varying reactances into the tuned circuit, which also changes oscillator operating frequency. Frequency stability, however, is the most important performance characteristic of an oscillator. This principal disadvantage may be eliminated only by using more complex, i.e., two-circuit or multicircuit oscillators.

Two-circuit oscillators. When a second tuned circuit is introduced into the oscillatory system of the oscillator, the principal possibility arises of dividing between them the functions listed above, namely, one of the circuits may be used for frequency stabilization, whereas the other may serve as an output load and also to provide coupling to an external load. The most simple way is to connect in the anode circuit of any self-oscillator discussed above an additional tuned circuit in series with the main one (Fig. 10-26). If the equivalent resistance of the additional circuit 2 is much greater than that of circuit 1 which determines the operating frequency, the greater fraction of the power being generated will be developed in circuit 2 and, therefore, the latter can be successfully coupled to an external load. The first circuit, on the contrary, is isolated as much as possible from all external effects (it may be either shielded or in some instances placed in a thermostat). External effects, however, may be transmitted from circuit 2 to circuit 1 through the valve anode-to-grid capacitance when triodes are employed in the oscillator.

To eliminate this capacitive coupling, the oscillators of this type employ tetrodes or pentodes (Fig. 10-26*b*) whose

screen grid grounded for high frequencies practically destroys the capacitive coupling between the tuned circuits. The only kind of coupling which cannot be removed in this oscillator is the coupling through the common electron stream passing between the electrodes to which the tuned circuits are connected. Therefore, this circuit configuration is known as an *electron-coupled oscillator*. As the screen grid and the upper point of circuit 1 are grounded (for high frequencies),

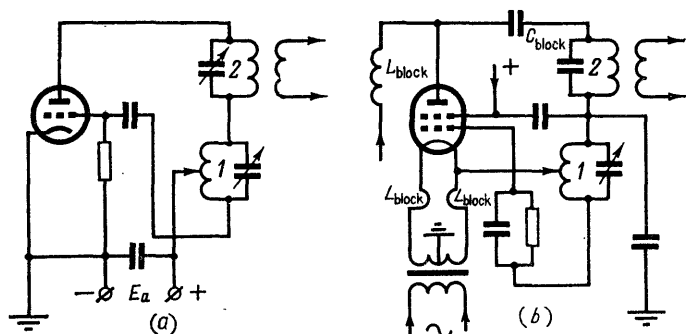


Fig. 10-26. Electron-coupled self-oscillator (electron-coupled oscillator)
(a) tuned circuit determining the oscillator frequency is isolated from a load with the aid of additional tuned circuit; (b) principle circuit

the cathode is under high-frequency voltage and cannot be directly grounded. Therefore, high-frequency blocking chokes L_{block} are generally connected in the filament circuit of directly heated valves.

Of particular interest is the possibility of obtaining in such an oscillator an increased output frequency by tuning circuit 2 to one of the higher anode-current harmonics. As this takes place, the mutual influence is still further weakened owing to the tuning of both circuits to different frequencies, which improves the frequency stability of the oscillator. The frequency limitations inherent in the single-circuit self-oscillators discussed above remain valid for this two-circuit oscillator as well.

Two-circuit self-oscillators in which tuned circuits are coupled through one of the valve interelectrode capacitances have found extensive application in the short- and ultrashort-

wave ranges. Here only two of the three reactances X_1 , X_2 and X_3 incorporated in the schematic diagram of a three-point self-oscillator represent circuit equivalent reactances whereas the third one is the reactance of the interelectrode capacitance that couples the tuned circuits.

Three oscillator configurations can be built on this principle (Fig. 10-27). In the oscillator whose schematic diagram is shown in Fig. 10-27a the tuned circuits are coupled through interelectrode capacitance C_{ag} . Both circuits have a

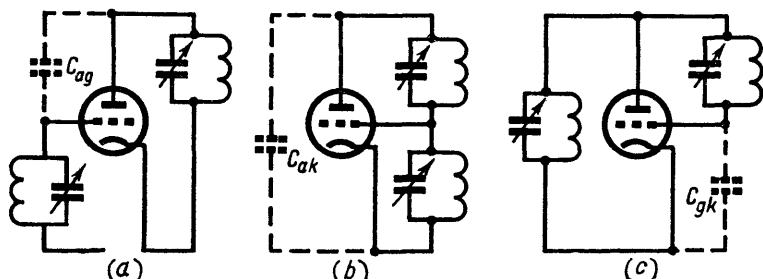


Fig. 10-27. Self-oscillators with tuned circuits coupled through interelectrode capacitances

(a) common-cathode oscillator; (b) common-grid oscillator; (c) common-anode oscillator

common point connected to the cathode. Therefore, this circuit configuration may be called a *common-cathode oscillator*. The common point is usually grounded so as to ground the rotors of the capacitors in both tuned circuits. In this case the circuit configuration is referred to as a *grounded-cathode oscillator*.

In the oscillator shown in Fig. 10-27b the tuned circuits are coupled through interelectrode capacitance C_{ak} . By the same indication as in the first case this circuit configuration is known as a *common (grounded)-grid oscillator*.

Finally, the tuned circuits in the oscillator illustrated in Fig. 10-27c are coupled through interelectrode capacitance C_{gk} . This circuit is called a *common (grounded)-anode oscillator*.

We shall discuss the performance characteristics of such oscillators in more detail since they are often used in radio systems.

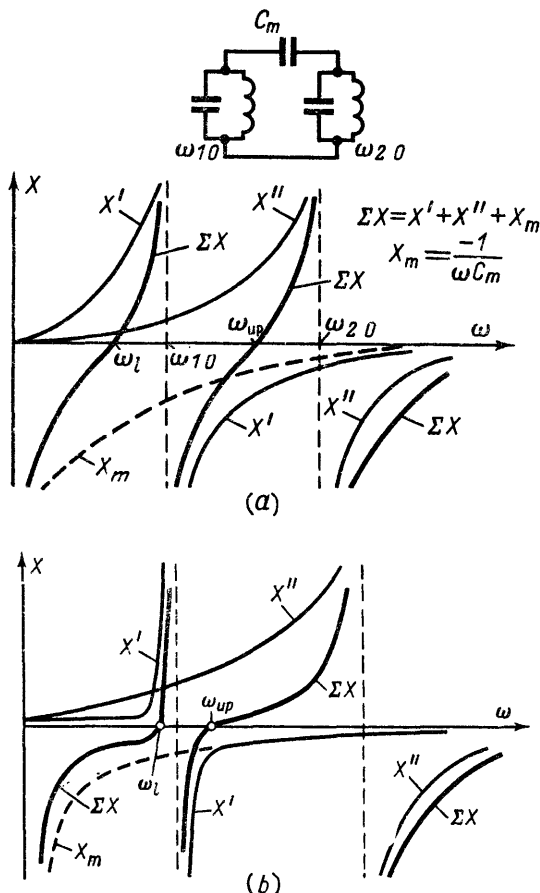


Fig. 10-28. Determining the natural frequencies of oscillatory system consisting of two capacitively coupled circuits

(a) quality factors of both circuits are identical; (b) quality factor of one circuit is much greater than that of the other

The system of two coupled circuits described in Chapter III possesses two natural frequencies (also called coupling frequencies), namely, the upper and lower frequencies. We find these frequencies by solving graphically equation (10-34). For this purpose, all three reactances are plotted on one graph

as functions of frequency and then summed up. It was shown in Chapter III that a parallel-circuit reactance tends to infinity with opposite signs on either side of the resonance frequency (see Fig. 2-40). In turn, the reactance of the coupling capacitance decreases monotonically with frequency. Assume that one circuit is tuned to frequency ω_{10} and the second one to frequency ω_{20} . For this case, reactances of the oscillatory system components are shown in Fig. 10-28a as functions of frequency.

The behaviour of the system total reactance, X , with varying frequency can be easily explained. When $\omega \rightarrow 0$, the reactance of the coupling capacitance tending to infinity with a negative sign dominates in the system, whereas when $\omega \rightarrow \omega_{10}$, the first-circuit positive reactance approaching infinity plays the dominant role in the system. Consequently, in the range from 0 to ω_{10} (i.e., in the frequency range of the first tuned circuit) one may find a frequency ω_l at which the total reactance will pass through zero, thus determining the lower coupling frequency of the system. Proceeding from similar considerations we may conclude that the upper coupling frequency ω_{up} at which the total reactance will pass once again through zero lies between the natural frequencies of the two tuned circuits.

It might be well to point out that if the quality factor of one of the tuned circuits is much greater than that of the other, both coupling frequencies will be close to the natural frequency of the high-quality tuned circuit. Such a case is illustrated in Fig. 10-28b. This can be explained as follows. The higher the quality factor of the tuned circuit, the steeper and closer the resonance characteristic curves pass to the resonance frequency.

Consider now which of coupling frequencies may become excited in two-circuit oscillators. For a lower coupling frequency lying below the natural frequencies of both tuned circuits their equivalent reactances will be inductive. For an upper coupling frequency lying between the natural frequencies of the tuned circuits their equivalent reactances will have opposite signs, i.e., the tuned circuit with the lower natural frequency will have a capacitive reactance, and the circuit tuned to the higher frequency will possess an inductive reactance.

In the common-cathode circuit reactances X_1 and X_2 are replaced by tuned circuits. The reactances of these circuits should have like signs in order to satisfy the phase condition of self-excitation. Consequently, an oscillator may get excited at the lower coupling frequency and not at the upper one. In this case, the oscillator becomes excited in compliance with the equivalent circuit of a tapped-coil oscillator.

In the common-grid configuration the capacitor C_{ah} plays the role of the reactance X_1 . Consequently, the tuned circuit used instead of the reactance X_2 should be equivalent to a capacitance. This is quite possible provided the coupling frequency at which self-excitation occurs lies above the natural frequency of the tuned circuit. In order for this to happen, the oscillator must become excited at the upper coupling frequency and the circuit connected in the place of X_2 (between the grid and cathode) must be tuned to the frequency which is less than that of the second tuned circuit inserted between the anode and grid. Thus, the oscillator is actuated in compliance with the equivalent circuit of the tapped-capacitor oscillator.

In the common-anode oscillator capacitance C_{gh} is used instead of reactance X_2 . Consequently, the equivalent resistance of tuned circuit X_1 should be capacitive. This is quite possible provided the oscillator gets excited at the upper coupling frequency and the circuit inserted between the anode and cathode is tuned to a much lower frequency than that connected between the anode and grid. Under these conditions the oscillator operates in compliance with the equivalent circuit of the tapped-capacitor oscillator.

Now, the question arises—how should the tuned circuits be adjusted in order to obtain the desired separation of functions between them? In the common-cathode oscillator the greater part of power is given off in the tuned circuit connected between the anode and cathode. Therefore, it should be coupled to an external load. In this case, the oscillator should be adjusted so that its frequency will be determined primarily by the other tuned circuit (connected between the cathode and grid) in which a low power is developed.

It is clear that the coupling frequency is affected basically by the circuit which is detuned to the least extent

relative to it. In this case, the coupling frequency lies in the region of abrupt changes in the equivalent reactance, $X_{eq}(\omega)$, of the given tuned circuit and, therefore, its retuning, as seen from expression (10-34), will change to a large extent the system total reactance which determines the oscillator frequency. This can also be seen from the graphs in Fig. 10-29 which illustrate variations in the oscillator frequency ($\omega'_l - \omega_l$) produced by strongly (Fig. 10-29a) and weakly (Fig. 10-29b) detuned circuits; in both cases the resonance frequency is changed by the same amount, i.e., $\Delta\omega_{20} = \Delta\omega_{10}$.

Since the lower coupling frequency is excited in the common-cathode oscillator, the cathode-grid circuit should be tuned to a lower frequency as compared with that of the anode-cathode circuit. The greater the difference between the resonance frequencies of these circuits, the less the influence of the anode-cathode circuit on the oscillator frequency but the less the power developed in it.

In the common-grid oscillator it is also desirable for its frequency to be determined by the cathode-grid circuit since the major part of the power is developed in the anode-grid circuit and the latter has to be coupled to an external load. As it will be shown later (when discussing the operation of a common-grid amplifier), the cathode-grid circuit being shunted by a low valve input impedance has a low quality and cannot provide a stable oscillation frequency. Thus, the desired separation of functions between the tuned circuits cannot be assured in the common-grid oscillator. As a consequence, this circuit configuration is not used for obtaining highly stable frequencies.

In the common-anode oscillator where the major part of the power is given off in the anode-cathode circuit and the upper coupling frequency is excited, it is desirable for the oscillator frequency to be determined by the anode-grid circuit. To accomplish this, the latter should be tuned to a higher frequency than the anode-cathode circuit. This is in agreement with the self-excitation conditions stated above.

Thus, we made sure that two of the three two-circuit oscillator configurations allow one to solve, to a great extent, the problem of separating the functions between the tuned

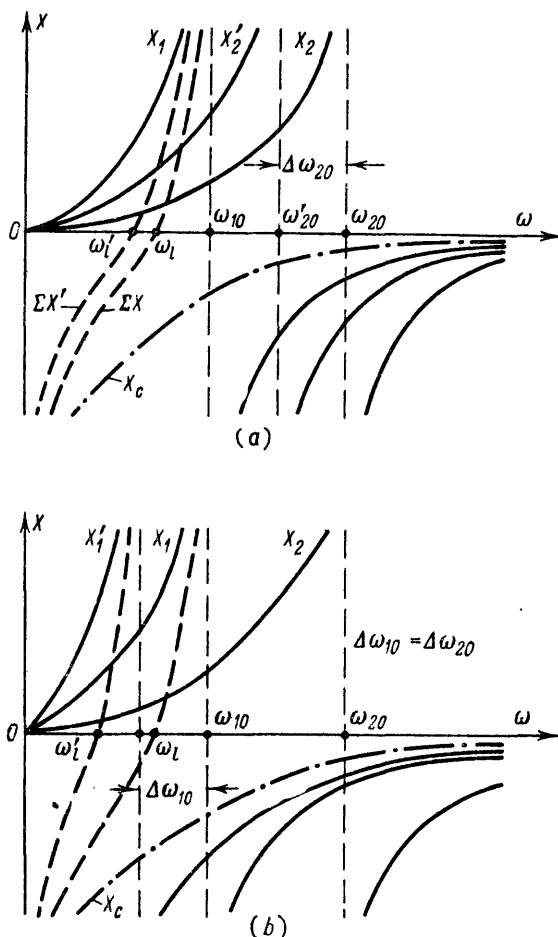


Fig. 10-29. Effect of retuning of one circuit to lower coupling frequency of system consisting of two capacitive-coupled circuits

(a) retuning of circuit with higher resonance frequency; (b) retuning of circuit with lower resonance frequency

circuits and obtain much more stable frequencies than those provided by single-circuit self-oscillators.

After discussion of the self-excitation process we may now

direct our attention to a very important problem of gain stability in separately excited oscillators.

Comparison between the separately excited oscillator (see Fig. 10-6a) and the common-cathode self-oscillator (see Fig. 10-27a) taking into account the presence of valve interelectrode capacitances shows that they do not differ from each other. This means that self-oscillations at the frequency satisfying equation (10-34) may also arise in a separately

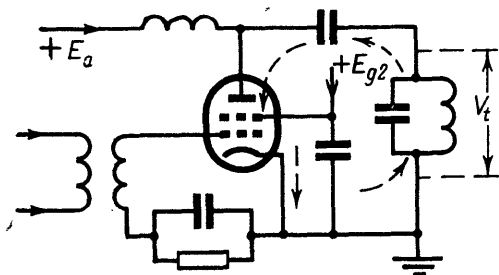


Fig. 10-30. Eliminating the stray capacitive coupling between anode and grid circuits with the use of tetrodes

excited oscillator provided the following conditions are satisfied: high coupling capacitance C_{ag} between the anode and grid circuits, high gain, and a proper difference between natural frequencies of the two circuits (i.e., their mutual detuning).

It is obvious that the probability of self-oscillations will be the greater the higher the oscillator gain since the less will be the feedback voltage (current) required to obtain amplitude balance. Therefore, sudden disturbances of normal operation are sometimes observed in a separately excited oscillator as its gain is increased and it transfers to the self-oscillation mode, or, in other words, the oscillator becomes unstable.

Feedback through valve interelectrode capacitances should be, of course, reduced in every possible way to eliminate this phenomenon. To solve this problem a four-electrode valve (tetrode) is employed which has a screen grid between the anode and control (first) grid. The screen grid is coupled for high frequency to the cathode (Fig. 10-30) through a

capacitor having a rather high capacitance. The screen grid shields the anode circuit from the grid one so that the high-frequency currents from the tuned circuit flow to the cathode through this grid, thus by-passing the control grid circuit.

Oscillators employing modern tetrodes provide stable amplification at wavelengths down to 30-20 cm. In high power oscillators operating at shorter wavelengths tetrodes are not used because of a number of technological and design difficulties. To obtain stable amplification in triode oscillators

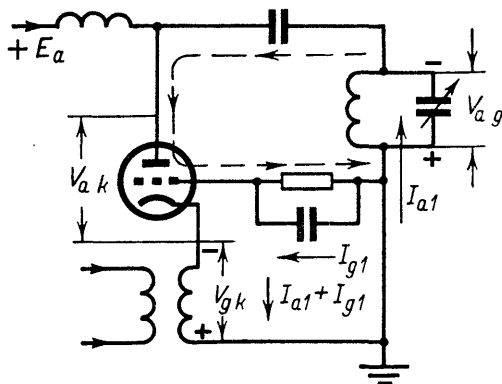


Fig. 10-31. Common-grid triode amplifier

capable of operating at these wavelengths, M.A. Bonch-Bruyevich suggested that the valve should be connected in the common-grid configuration (Fig. 10-31) in which the control grid functions as a screen separating the input and output circuits. In such an oscillator feedback is accomplished through capacitance C_{ak} which dozes and hundreds of times less than capacitance C_{ag} . Common-grid amplifiers have found extensive application in the short- and ultra-short-wave (microwave) ranges.

Let us consider the major features of the common-grid amplifier. As distinct from the common-cathode amplifier, here through the input circuit flows not only grid current but the anode current as well which is many times as great as the former. Therefore, the exciting power for this amplifier

defined as

$$P_{ex} = \frac{1}{2}(I_{a1} + I_{g1})V_{gk} = \frac{1}{2}I_{a1}V_{gk} + \frac{1}{2}I_{g1}V_{gk} \quad (10-41)$$

will be much greater than that required for the common-cathode amplifier. The first term in the right-hand side of this expression represents an additional exciting power as compared with that of the common-cathode oscillator.

One should not think that this power is uselessly lost. By Kirchhoff's law and using the designations of Fig. 10-31 we may write:

$$V_{gk} + V_{ag} + V_{ak} = 0 \quad (10-42)$$

The signs of these voltages depend upon the choice of a reference point. If voltage V_{ag} is taken to be the voltage drop across the output circuit, i.e.,

$$V_{ag} = I_{a1}R_{eq} \quad (10-43)$$

it will rise when voltage V_{gk} between the grid and the cathode is positive. Let us assume that the direction of anode current is positive. It can easily be seen that voltages V_{ag} and V_{gk} have opposite directions (see Fig. 10-31), i.e., they are 180° out of phase. Thus, from equation (10-42) we obtain

$$V_{ak} = V_{ag} - V_{gk}$$

or

$$V_{ag} = V_{ak} + V_{gk} \quad (10-44)$$

Power developed in the anode-grid circuit is

$$P = \frac{1}{2}I_{a1}V_{ag} = \frac{1}{2}I_{a1}V_{ak} + \frac{1}{2}I_{a1}V_{gk} \quad (10-45)$$

The first term in the right-hand side of (10-45) represents the power delivered by the valve to the output circuit, the second term determines the additional power drawn by the amplifier from the exciter. Consequently, this additional power is transmitted from the exciter to the amplifier output.

High exciting power decreases appreciably the amplifier power gain which usually ranges from 6 to 20. A current increase in the input circuit of the amplifier reduces its input impedance determined by

$$R_{in} = \frac{V_{in}}{I_{in}} = \frac{V_{gk}}{I_{a1} + I_{g1}} \quad (10-46)$$

Here, R_{in} amounts to dozens and hundreds of ohms as compared with thousands of ohms in the common-cathode amplifier. The last feature is useful in wideband amplifiers. Moreover, the low input impedance of this amplifier makes it possible to feed its input circuit through low-resistance coaxial feeders.

10-4. ULTRAHIGH-FREQUENCY VALVE OSCILLATORS

At metre wavelengths and shorter some specific effects appear in oscillators, which call for changes in design of both valves and oscillatory systems.

At higher frequencies the reactance of valve interelectrode capacitances increases. For instance, when changing the wavelength λ from 100 m to 1 m or 10 cm, capacitance reactances decrease 100 and 1000 times, respectively. Therefore, at the same voltages applied to the electrodes currents in valve leads rise accordingly. In addition, heat losses in valve leads rise proportionally to the square of currents and increase still more with increasing frequency due to the skin effect. Overheating of the leads in early ultrahigh-frequency valves generally resulted in the envelope cracking at lead seals. To eliminate overheating, the leads of such valves are made in the form of thick rods, wide ribbons or disks. Simultaneously, steps are taken to reduce valve interelectrode capacitances.

Furthermore, at higher frequencies dielectric losses increase significantly since polarization of molecules in dielectrics alters at a greater rate. The regions of joints between electrode leads and envelope glass (metal-glass seals) undergo high-frequency heating under the action of the applied voltages. Glass losses are rather high, and glass envelopes are very critical to temperature variations. Therefore, special measures are generally taken to avoid the softening of envelope glass, to retain valve ruggedness and reduce losses. Modern UHF valves are usually enclosed in ceramic envelopes; this material has high thermal resistance and low losses.

The inductance of valve leads increases with rising frequency and, therefore, grids cannot be used as screens at high frequencies. Moreover, at higher frequencies lead inductance presents an ever increasing resistance to capacitive

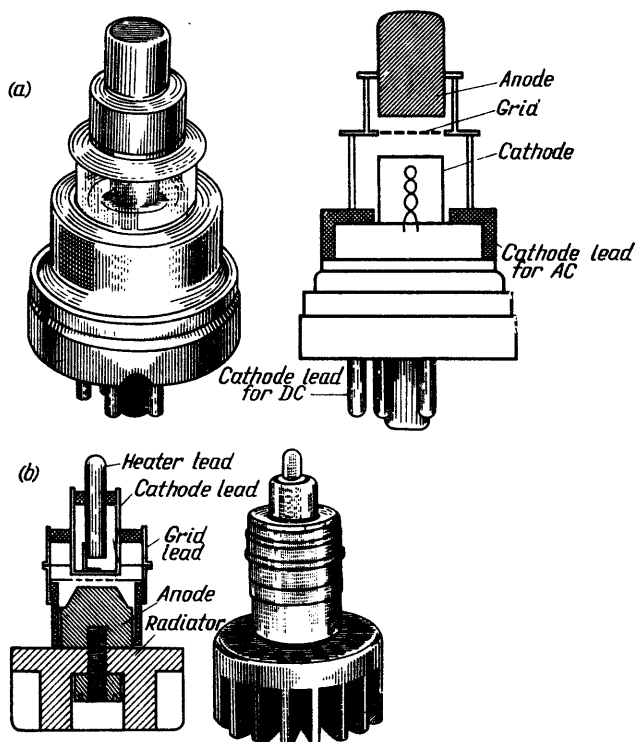


Fig. 10-32. Typical constructions of UHF oscillator triodes
 (a) low-power lighthouse triode with disk leads (disk-seal valve); (b) high-power metal-ceramic triode with cylindrical leads

currents flowing from the output circuit through the screen grid in the common-cathode circuit configuration (see Fig. 10-30) and through the control grid in the common-grid circuit configuration (see Fig. 10-31). As a consequence, the feedback between the input and output circuits increases so that attainable stable gain reduces. To reduce lead inductance, planar electrodes with disc and cylindrical leads are used instead of spiral wire electrodes.

The typical constructions of UHF triodes with planar electrodes, disc and cylindrical leads are illustrated in Fig. 10-32. Some of the specific features of these designs will be more clear after discussion of UHF oscillatory systems.

In the tuned circuits designed for higher frequencies, components with lower inductances and capacitances are required. This means that in the metre wave range capacitors will have only two or three pair of small plates and inductors, only a few turns. In a high-frequency tuned circuit the designer attempts to reduce capacitance in every possible way so that inductance and, consequently, characteristic impedance can be increased since

$$\rho = \sqrt{\frac{L}{C}}$$

This is important since a high load resistance, i.e., the equivalent resistance of tuned circuit $R_{eq} = Q\rho$ required for valve oscillators is provided. This is why external capacitors are often discarded even in the metre wave range and only valve interelectrode capacitances are used instead. In this case, tuning is accomplished by changing circuit inductances. Constructionally, this is most conveniently achieved with the use of short-circuited segments of long lines as inductances.

As shown in Chapter V, short-circuited lines of a length less than $\lambda/4$ have an inductive input impedance changing from zero to infinity with wavelength by the following law:

$$X_{in} = Z_0 \tan \frac{2\pi l}{\lambda}$$

If capacitor C is connected to the line input, resonance will be obtained provided the following condition is satisfied:

$$\frac{1}{\omega C} = Z_0 \tan \frac{2\pi l}{\lambda}$$

hence, the line resonance length will be

$$l_1 = \frac{\lambda}{2\pi} \tan^{-1} \frac{1}{\omega C Z_0} \quad (10-47)$$

As seen from the graphs of Fig. 10-33, the line resonance length decreases with the increasing capacitance. The same figure shows that owing to the periodic variations of line impedance as line electrical length changes, resonance will occur not only at length l_1 but also when

$$l_2 = l_1 + \frac{\lambda}{2}; \quad l_3 = l_1 + 2\frac{\lambda}{2}; \quad l_4 = l_1 + 3\frac{\lambda}{2} \text{ etc.}$$

This property is useful when l_1 is too small and the line becomes too difficult to construct, then the line length is increased by an integral number of half-waves.

Segments of balanced two-wire lines are most frequently employed in push-pull oscillators which have found exten-

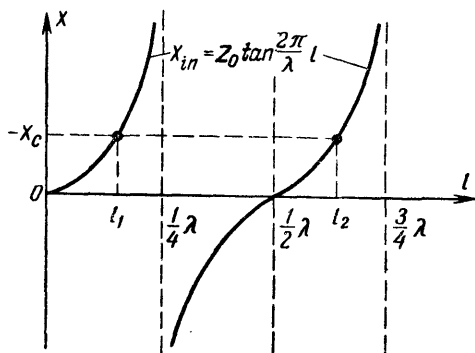


Fig. 10-33. Determining the resonance length of a segment of short-circuited line with capacitor connected at its input

sive application in the short and ultrashort wave ranges. Such oscillators employ two identical valves each connected to the load from opposite sides as shown in Fig. 10-34. Exciting voltage is applied to the valve grids in antiphase and, therefore, currents in both circuit halves (arms) are also in antiphase. If anode current in one of the valves is

$$i'_a = I'_{a0} + I'_{a1} \cos \omega t + I'_{a2} \cos 2\omega t + I'_{a3} \cos 3\omega t + \dots$$

anode current in the second valve has the same amplitude but is 180° out of phase with the exciting voltage, i.e.,

$$i''_a = I''_{a0} + I''_{a1} \cos \omega \left(t + \frac{T}{2} \right) + I''_{a2} \cos 2\omega \left(t + \frac{T}{2} \right) + I''_{a3} \cos 3\omega \left(t + \frac{T}{2} \right) + \dots$$

Opening the brackets and using periodicity of trigonometrical functions, we obtain

$$i''_a = I''_{a0} - I''_{a1} \cos \omega t + I''_{a2} \cos 2\omega t - I''_{a3} \cos 3\omega t + \dots$$

Currents flow in the opposite directions through a common load. If the circuit halves are balanced, the resultant current through the load will be

$$i_{a,r} = i'_a - i''_a = 2I_{a1} \cos \omega t + 2I_{a3} \cos 3\omega t + \dots \quad (10-48)$$

In common supply networks, on the contrary, currents flow in one direction and the resultant current is determined as

$$i_{a,r0} = i'_a + i''_a = 2I_{a0} + 2I_{a2} \cos 2\omega t + \dots \quad (10-49)$$

Expression (10-48) indicates that the fundamental and all uneven harmonics act on the load in phase, thus doubling the power of the oscillations being generated. The even harmonics act 180° out of phase and are mutually cancelled. From expression (10-49) it follows that the processes occurring in common supply networks are just opposite since here the direct components and all even harmonics are summed up and all uneven harmonics are cancelled out.

One of the most important features of the push-pull circuit configuration is in that valve interelectrode capacitances

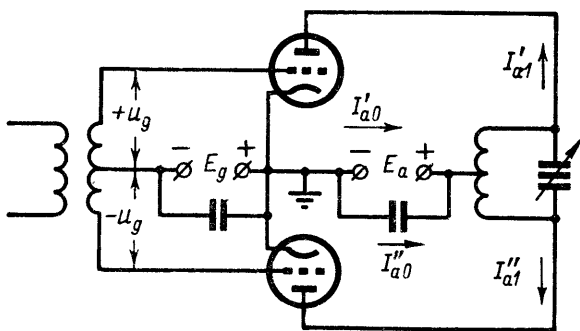


Fig. 10-34. Push-pull connection of oscillator valves

when being involved in the tuned circuit prove to be connected in series so that their resultant capacitance is reduced twice. It has already been mentioned how important it is to reduce circuit capacitance in the UHF range. It is for this reason that push-pull oscillators are widely used in this frequency range.

It is convenient to use two-wire lines in push-pull circuits not only because of their symmetry, but owing to the fact that voltages across their wires are always 180° out of phase relative to the ground, which is important for push-pull oscillators.

The schematic diagram shown in Fig. 10-35 is an example of a separately excited push-pull oscillator employing the common-grid circuit configuration. In this oscillator tuned circuits are formed by the segments of a two-wire line and interelectrode capacitances. As the wires are under the opposite potentials relative to the ground, the centre points of short jumpers always have the ground potential for high frequencies so that it is convenient to connect a supply source to these points.

Two-wire lines start, rather intensively, to radiate energy into space, just like antiphased antennas do (see Chapter VI), when the distance between their wires $d \geq 0.1\lambda$. The radiation may be eliminated by using coaxial-line segments in which the field is enclosed between the inner and outer conductors (tubes). Such oscillatory systems are widely applied in the decimetre waverange.

For convenience of direct connection of coaxial lines to valves, the electrode leads of the latter are made in the form of short cylinders each having different diameters (see Fig. 10-32) into which three coaxial tubes are plugged in (Fig.

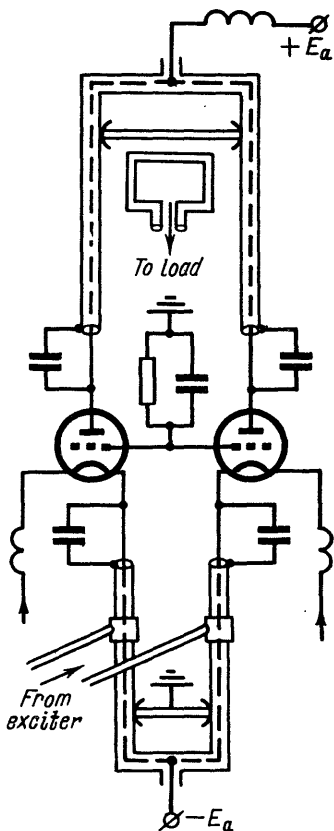


Fig. 10-35. Separately excited push-pull oscillator operating in metre band

10-36). The inner surface of the outer tube and the outer surface of the intermediate (second) tube form an anode-grid cavity connected to interelectrode capacitance C_{ag} . The line length and, consequently, the resonance frequency of the cavity are changed with the aid of a movable short-circuiting plunger. The inner surface of the intermediate tube and the outer surface of the central (small) tube form the cathode-

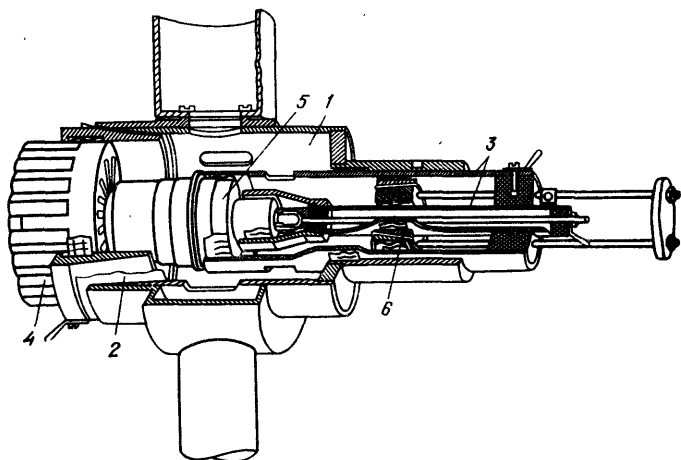


Fig. 10-36. Structure of decimetre-band oscillator employing metal-ceramic triode connected in common-grid circuit

1 — anode-grid cavity; 2 — insulating gasket; 3 — cathode-grid cavity; — radiator; 5 — metal-ceramic triode; 6 — short-circuiting plunger

grid cavity connected to interelectrode capacitance C_{gh} . This cavity is tuned in the same way as the preceding one. In such oscillators coupling with coaxial cavities is most frequently accomplished through a coupling loop inserted into the region where the magnetic field strength is a maximum (near the short-circuiting plunger) or by means of a capacitor in the form of a small plate placed into the region with the maximum electric-field strength (near the valve electrodes).

Improvements in electron valves and oscillatory systems designs brought about in the last few years made it possible to build valve oscillators that can operate at waves up to 30 cm as successfully as those employed at much longer

waves. In the near future it is expected that such oscillators will be used in the entire range of decimetre waves. Some low-power triodes are available even now which can operate in the entire centimetre waverange though they provide rather low efficiency.

All the above concerning separately excited oscillators refers to UHF self-oscillators as well. The latter employ the same circuit configurations with the same design ele-

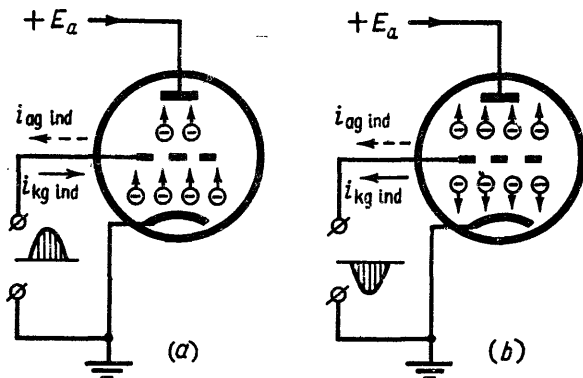


Fig. 10-37. Currents induced in grid circuit by electrons; transit time is comparable with oscillation period

ments. The application of metal-ceramic valves (disk-seal valves) with the above arrangement of electrode leads calls for the common-grid circuit configuration of the oscillator.

Additional difficulties are associated with the fact that the valve no longer functions as an inertialess current regulator at centimetre waves. The transit time, i.e., the time required for electrons to travel between two electrodes in an electron valve becomes comparable with the oscillation period; this has the following consequences.

It was shown in Section 10-2 that a uniform electron flow is controlled without power losses when as many electrons reach the valve grid as leave it. When the transit time becomes comparable with the oscillation period, the electron flow is no longer uniform. During a positive half-cycle of exciting voltage (Fig. 10-37a) the number of electrons above the grid will be less than that below it since not all electrons

have a chance to pass through the grid in the given period. As a consequence, the resultant induced current ($i_{hg\ ind} - i_{ag\ ind}$) appears in the grid circuit. During a negative half-cycle (Fig. 10-37*b*) the electrons which have not reached the anode or the grid by this moment will move away from the grid, thus causing the appearance of induced current ($i_{hg\ ind} + i_{ag\ ind}$). All this requires an increase in exciting power and, consequently, worsens the amplifying properties of the oscillator.

But the main thing resides in the fact that a finite transit time leads to a delay of electrons with respect to the voltage phase which is the most favourable for energy transfer to the tuned circuit and, therefore, some electrons fall within the unfavourable phase. As a consequence, oscillator electron efficiency and power are reduced. The last factor limits the application of valve oscillators in the range of centimetre waves but even there they are replaced by oscillators based on other principles.

10-5. KLYSTRON AMPLIFIERS AND OSCILLATORS

As already stated, the finite transit time results in low efficiency of valve oscillators in the UHF. However, it appears that electron inertia may be used to produce short current pulses which are necessary for obtaining high electron efficiencies of oscillators. A two-cavity klystron was the first device in which this principle was realized.

The structure of such a klystron is shown in Fig. 10-38. In the left-hand end of the envelope is an electron gun whose purpose is to produce a uniform flow of high-velocity electrons. It consists of an oxide-coated cathode 1, focusing electrode 2 made in the form of a diaphragm or cylinder supplied with a low negative voltage to compress the electron beam transversely and an accelerating electrode 3 which is under a high positive voltage. On leaving the electron gun, the uniform flow of electrons moves farther by inertia since the remaining elements of the klystron have the same potential as the accelerating electrode and there are no constant electric fields after the electron gun. A cavity resonator whose centre portion is made in the form of grids is arranged on the path of the electron beam.

Cavity resonators, just like the coaxial cavities discussed above, are formed by closed metal surfaces and represent oscillatory systems with natural frequencies lying in the

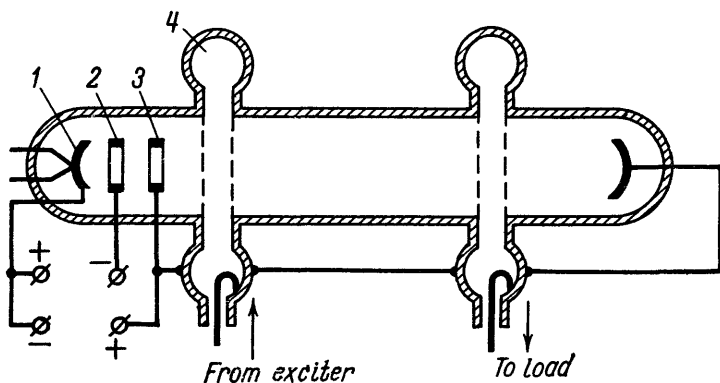


Fig. 10-38. Two-cavity klystron used as amplifier

range of decimetre, centimetre and millimetre waves. Their origin may be explained as follows.

In order to obtain a tuned circuit with the highest possible natural frequency, a designer endeavours to reduce its

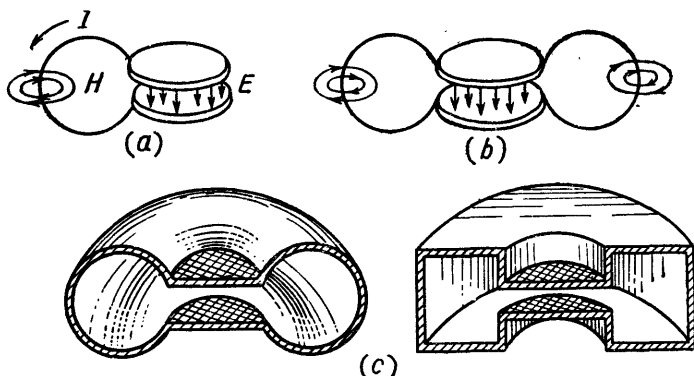


Fig. 10-39. Cavity resonators

inductance and capacitance as much as possible. While doing so, he reduces its capacitor to two small plates and the

inductor to one turn connecting these plates (Fig. 10-39a). At first sight it seems that nothing more can be done. However, we know that the resultant inductance is decreased in the case of a parallel connection. Now, it is quite evident that it is advisable to connect in this way as many turns as possible (Fig. 10-39b). By infinitely increasing the number of turns, we shall obtain the continuous metal surface

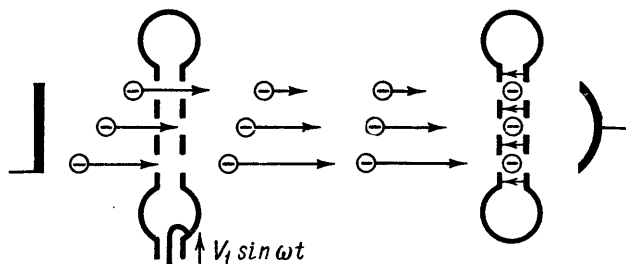


Fig. 10-40. Speed modulation and bunching of electron stream in klystron

illustrated in Fig. 10-39c. The natural frequency of such a "tuned circuit" will be extremely high.

Incidentally, it should be noted that cavity resonators display a high quality factor which amounts to hundreds and even thousands owing to the large surface of current-carrying elements and the absence of radiation losses (since the entire field is enclosed within the cavity).

The first cavity resonator (the first tuned circuit) of the klystron is fed (through a feeder line) with voltage $V_1 \sin \omega t$ at a frequency equal to its natural frequency (Fig. 10-40) to excite forced oscillations in the cavity. Electrons which have entered into the space between the resonator grids are subjected to the influence of the resonator electric field. Some of these electrons are retarded and others accelerated. Since oscillations in the first resonator are weak, changes in electron velocity are relatively small and they do not cause flow density changes inside the resonator.

While moving farther beyond the resonator, the accelerated electrons begin overtaking the electrons which have passed the resonator when voltage across its grid became zero after a retarding phase (these electrons are called *undisturbed* electrons). Electrons which have passed the resonator ear-

lier in the retarding field will lag behind and approach also the undisturbed electrons, and at some distance from the first resonator, called the *buncher*, the electrons will bunch near the undisturbed electrons.

It should be noted that the reverse will occur with electrons which enter the cavity resonator when the resonator field is changing its phase from the accelerating to the retarding. Since the cavity resonator is passed first by high-velocity and then by low-velocity electrons they will both move away from the undisturbed electrons so that rarefied gaps will form near them in the electron flow. Therefore, the bunches of electrons will be spaced by intervals equal to the oscillation period in the buncher. Such a process when low-velocity modulation leads to electron stream density modulation is called *velocity-modulation of electron stream*. The space next to the buncher where this process occurs is known as the *drift space* since electrons move (or drift) in it by inertia.

If we place the second cavity at the gap where the bunches of electrons are formed, the electrons will induce in this cavity alternating currents as they pass through its grids, and an alternating electric field will be set up between these grids. The bunches of electrons pass once per cycle, and by interacting with the cavity resonator retarding field replenish resonator energy just like anode current pulses replenish tuned circuit energy in a valve oscillator. In much the same manner as in the valve oscillator (see Section 10-1) power expended on the velocity modulation of a uniform electron flow is practically negligible, this means that the device possesses high amplification.

The second cavity which extracts energy from the bunched electron stream is called the *catcher*. As the bunched stream acts on the catcher just like the sea waves, the device is called the *klystron*. (The word "klystron" is derived from the Greek "klyzein" which refers to breaking of waves on a beach.) Electrons that have passed the catcher are gathered by the collector. Useful RF power is extracted from the catcher by a coupling device and delivered to a load through a feeder line. As any amplifying device, the klystron may be converted into a self-oscillator if the catcher is coupled to the buncher through a feedback path (Fig. 10-41).

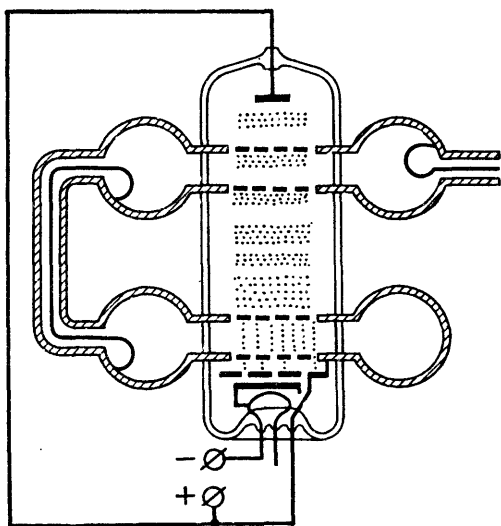


Fig. 10-41. Two-cavity klystron used as self-oscillator

Two-cavity klystrons and more complex multicavity klystrons operating on the same principle are used extensively as amplifiers at decimetre and rarely centimetre waves. Their chief advantage is a high power gain amounting up to dozens of millions. Among disadvantages of klystrons are their comparatively low efficiency, not exceeding 45 per cent, and a very narrow frequency range determined by the presence of two (or more) high-quality resonators which can hardly be tuned mechanically.

Single-cavity *reflex klystrons* are widely used as low-power self-oscillators. In this klystron a repelling electrode (reflector) having a negative voltage is used instead of the second cavity resonator (Fig. 10-42). Electrons which have passed the resonator move by inertia to the reflector but come to a stop under the action of its field and then pass through the resonator once again but this time in the reverse direction. Accelerated electrons come closer to the reflector than retarded electrons. Reflector voltage may be adjusted so that both groups of electrons will simultaneously return to the resonator. The bunches of electrons thus formed give

up more energy to the cavity resonator than it spends on changing the velocity of the uniform electron flow.

Oscillations in the klystron are initiated in the same way as in any other self-oscillator. When the device is switched on or any other disturbance occurs, the primary current pulse sets up weak oscillations in the resonator. These oscillations modulate the velocity of the electron stream,

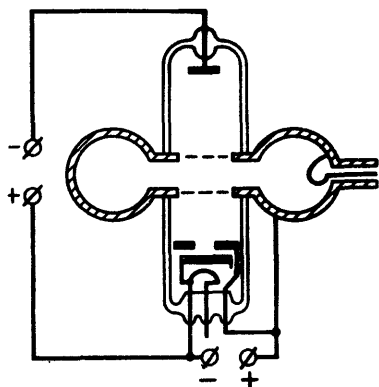


Fig. 10-42. Single-cavity reflex klystron

thus the bunches of electrons are formed which act on the cavity resonator at the modulation frequency. The oscillations increase in the resonator owing to the klystron amplifying properties mentioned above. Oscillation amplitude gradually becomes constant as a result of increasing resonator losses and also due to the fact that with a high resonator voltage the velocity modulation of the electron stream increases, electrons will return to the ca-

vity resonator at different moments, thus deteriorating the electron stream bunching.

The distinguishing feature of reflex klystrons is the possibility of changing the oscillation frequency electronically, i.e., by changing the voltage across the reflector. Let us explain this process physically. Assume that the bunches of returning electrons i pass through the cavity resonator during the maximum retarding field phase (Fig. 10-43a). The interaction between this field and the electron flow is of a purely active nature. This means that the first harmonic of the electron-beam current i_1 passes through the resonator in phase with the voltage u_{rnt} applied to it. If we change reflector voltage, the time when the bunches of electrons reach the resonator will change too. Under these conditions they will not pass through the resonator precisely in the maximum retarding phase (Fig. 10-43b). As a result, the first harmonic of the electron beam current will be shifted

accordingly in time. Therefore, a phase shift will appear between the first harmonic and the resonator voltage, i.e., reactive interaction will take place in addition to the active one. The reactive interaction will cause an appropriate change in the resonator natural frequency which, in turn, leads to changes in the oscillation frequency until the interaction becomes active. Such a phenomenon is known

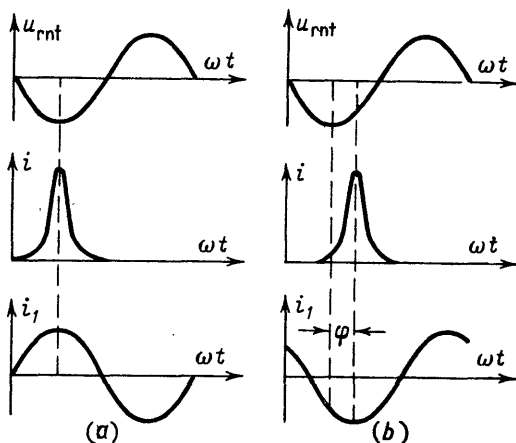


Fig. 10-43. Electronic tuning in reflex klystron

as *oscillator electronic tuning*. In reflex klystrons the relative range of electronic tuning does not exceed usually several per cent.

The resonators have to be reduced in size with the shortening wavelength. Simultaneously, the voltage across the resonator should be also reduced in order to prevent electrical breakdown. As a consequence, the generated power decreases. In addition, the time required for electrons to travel between the resonator grids becomes comparable with the oscillation period; this leads to the difficulties we have already encountered in valve oscillators, i.e., to the necessity of increasing the driver power, low efficiency of interaction between the electron stream and the field, and to low gain. A search for other methods of interaction between the electron stream

and the field produced by an oscillatory system has culminated in the advent of a wide variety of travelling-wave oscillators.

10-6. TRAVELLING-WAVE OSCILLATORS

In a klystron electrons are exposed to relatively intensive field during short time intervals in order to obtain velocity modulation. The net effect will be about the same, when the electrons are acted upon by a much weaker field provided the time of its action is increased accordingly. The same may be said concerning the interaction between the bunched electron stream and the field of an output oscillatory system.

These considerations form the foundation for development of high power oscillators generating very short waves as well as oscillators with a wide range of electronic tuning. The last problem cannot be solved with the aid of klystrons since their oscillatory system involves high-Q narrow-band resonators. Obviously, the application of wideband or aperiodic (non-resonant) electromagnetic systems can help to resolve this problem. However, these systems both produce much weaker field than those formed of high-Q resonators.

As is known from Chapter V, a matched transmission line along which travelling waves are propagated may serve as an ideal wideband aperiodic system. Let us assume that a travelling wave is produced in such a line by a small signal applied to its input. If an electron stream moves along the line in the same direction and at the same velocity as the travelling wave, then some electrons will move in a retarding field, during the entire travel, and others in an accelerating field. This will lead to velocity modulation of the electron stream so that at some distance from the line input all the electrons, just like in the klystron, will bunch near the undisturbed electrons.

However, under these conditions no energy can be transferred to the wave by electrons since the electron bunches move all the time in the wave zero field (Fig. 10-44a). The situation will be quite different if the electron-stream velocity v_e slightly exceeds the wave phase velocity v_w . In this case, as before, some electrons will move for a relatively long period of time in the retarding field and others in the

accelerating field since the difference between the two velocities is very small. As a result, the velocity modulation will occur and the electrons will be grouped near undisturbed electrons. These electron bunches gradually leading the wave will fall within its retarding field (Fig. 10-44*b*). And again due to the small difference in the velocities they will move for a considerable period of time in the retarding field of

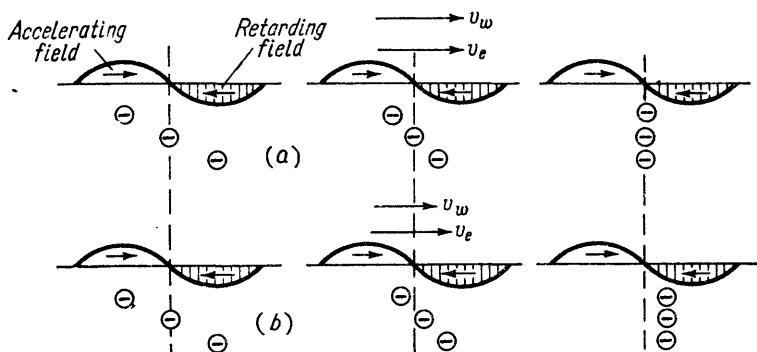


Fig. 10-44. Interaction between travelling wave and electron stream
(a) electron-stream and wave velocities are equal; (b) velocity of electron stream exceeds slightly that of wave

the wave, thus giving up their energy to it. When the electron bunches pass into the accelerating field, they will take away some energy from the wave. Therefore, for such a system some optimum length exists depending upon the difference between velocities of the electron stream and the wave and upon the frequency.

One may say that the idea is clear but some principal difficulties lie on the way of its realization. First, electromagnetic waves propagate along aerial lines practically at the velocity of light, whereas, according to the relativity theory, electrons cannot move with a greater speed. Second, as shown in Chapter V, the electric field of a transmission line is perpendicular to the direction of wave propagation and, therefore, it cannot interact with the electrons moving along the line wires.

All this is absolutely true. One may even add that if interaction could take place even at an electron velocity some-

what less than the velocity of light, severe problems associated with the use of voltages of hundreds of thousands of volts, required for imparting such a high velocity to the electrons, would arise. Consequently, to solve the problems indicated, one should use such transmission lines in which the velocity of electromagnetic waves propagation is much less than that of light, i.e., develop a system slowing down waves, and the wave electric field in these systems should

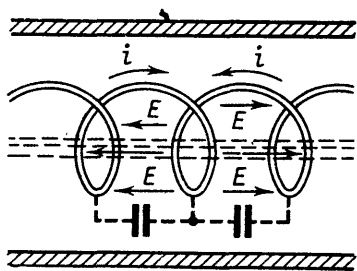


Fig. 10-45. Helical slow-wave structure (coaxial)

have a longitudinal component. Such systems are called slow-wave structures.

A coaxial line, for instance, whose centre conductor is wound into a helix (Fig. 10-45), may serve as a slow-wave structure. For the sake of simplicity in understanding the properties of such a structure, let us imagine that the wave travels around the helix turns.

Then, the velocity of wave propagation along the helix axis is slowed down (a "slow-down" of the order of ten is attainable). After running around a helix turn, the wave changes its phase. As a consequence, a potential difference is set up between the adjacent turns, thus causing the component directed along the helix axis to appear in the wave field. It is precisely this component that will act on the electron stream.

It is not difficult now to visualize the structure of a travelling-wave tube (Fig. 10-46). At the left end of the tube envelope an electron gun is located. The signal to be amplified is applied to the helix input through a waveguide or feeder. This signal excites a travelling wave in a slow-wave structure along which the electron beam moves. The velocity modulation occurs in the initial portion of the slow-wave structure and does not require high power since the electron stream is uniform. Then electrons are grouped and give up their energy to the wave whose amplitude increases many times on its way to the output. The length of the slow-wave structure is selected so that the electron bunches can pass the entire retarding half-wave before reaching the out-

put. The electrons that have passed the slow-wave structure are gathered by a collector.

Appreciable forces of electrical repulsion exist in the bunches of electrons. If the path of electron stream is long

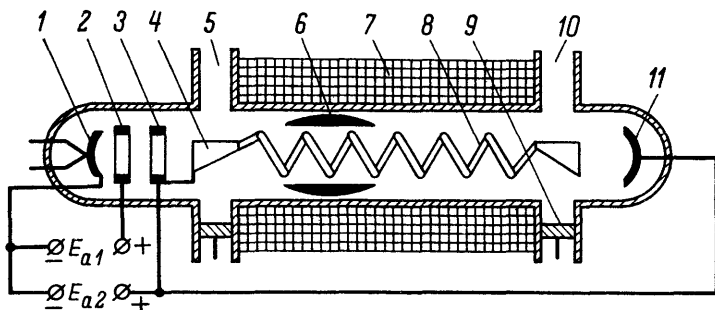


Fig. 10-46. Travelling-wave amplifier

1—cathode; 2—1st anode; 3—2nd anode; 4—conical wideband receiving dipole; 5—input waveguide; 6—absorber; 7—focusing coil; 8—helix; 9—matching stub; 10—output waveguide; 11—collector

enough, these forces may cause beam divergence and the settling of electrons on the helix. To prevent beam divergence, the tube is generally placed in the longitudinal magnetic field produced by a coil or permanent magnet. Assume that under the action of the repulsive forces an electron diverts from the helix axis. Then, it will fall also under the action of a deflecting force of the magnetic field which is perpendicular to the helix axis. As a result, the electron will change its direction (Fig. 10-47). If we trace in succession the path of an electron moving in the coil field, it turns out that the electron moves along a complicated helical trajectory which approaches the axis of the slow-wave structure. Consequently, the coil focuses the electron stream into a narrow beam moving along the axis of the tube.

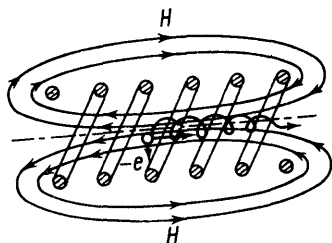


Fig. 10-47. Focusing action of solenoid longitudinal magnetic field

The amplifiers employing travelling-wave tubes (TWT) possess a high gain mounting up to hundreds of thousands of times. As any other amplifier, the travelling-wave tube may be used as a self-oscillator by coupling its output to input. This may be accomplished with the aid of a special feedback line or simply by producing some mismatch at the output of the line (the slow-wave structure), then some energy will be reflected back to the line input. If the reflected wave is sufficiently intensive and in phase with the wave at the line input, self-excitation occurs. In amplifiers where self-excitation should be eliminated but some mismatch is always available, a special absorber, designed to attenuate reflected waves, is usually placed at the centre of the slow-wave structure.

Owing to the absence of resonance systems, travelling-wave tubes used as amplifiers have a very wide passband with the lower-to-upper frequency ratio of 1:4 for some specimens. They are commonly used in the decimetre, centimetre and millimetre wave ranges. Travelling-wave tubes are employed rarely as self-oscillators since the above self-excitation condition requiring a definite phase of the reflected wave narrows heavily the operating range of the oscillator. Backward-wave tubes (BWT) open up much greater possibilities to wideband operation.

10-7. BACKWARD-WAVE OSCILLATORS

Apart from the helix, travelling-wave tubes may employ some other slow-wave structures. For instance, comb lines (Fig. 10-48a) in which, simply speaking, the slowdown is obtained due to the fact that the wave runs in succession around comb slots, have found extensive application. For the time the wave runs around a slot, voltage across the slot "input" changes, and between the slot edges a potential difference and longitudinal field interacting with the electron stream are set up (Fig. 10-48b). A coaxial line with the inner conductor of the shape shown in Fig. 10-48c exhibits similar properties. In the centimetre-wave range disk-loaded waveguides (Fig. 10-48d) are used instead of the coaxial line.

Assume, for instance, that an electromagnetic wave producing the maximum retarding field in slots 1, 3, 5, etc. at

the given instant is propagated from right to left along a comb-shaped structure. Let us put the question: can the electron stream interact effectively with the field of such a system as it travels toward the wave? In the helix slow-wave structure the interaction between the electron stream moving in one direction and the wave propagating in the opposite direction (the backward wave) cannot produce the desired effect since the electrons are alternately influenced by a retarding

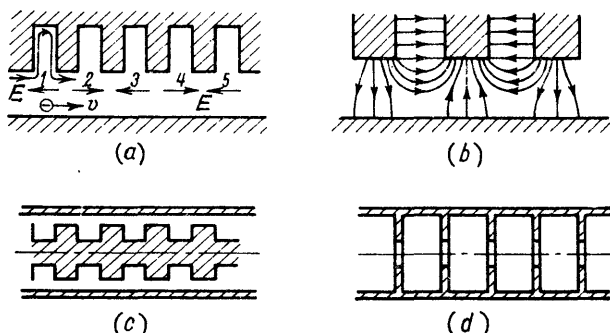


Fig. 10-48. Nonuniform slow-wave structures

field and an accelerating field. The structure of the field in the comb line, however, is different (Fig. 10-48b) since the maximum longitudinal field acts here between the slot edges, whereas no field exists under the slot tooth. Therefore, if the electron that has passed the first slot in the retarding field reaches the second slot after the field has changed by a half-cycle, it will be subjected to a retarding action in the second and all the succeeding slots. In doing so, the electron will pass under the comb teeth by inertia without undergoing the action of the backward wave, since the field longitudinal component is zero there at all times. Consequently, both the bunching of electron stream and the transfer of energy to the wave by the bunched electron stream are possible in such a slow-wave structure.

Periodic structures in the separate parts of which an alternating longitudinal electric field acts and in other parts does not are called *nonuniform* systems as distinct from the helices along which the electric field has the same amplitude and, therefore, the latter are referred to as *uniform* systems.

Electron devices in which interaction occurs between the electron stream and the backward wave propagating along a nonuniform slow-wave system are called *backward-wave tubes*.

In the example discussed above the effective interaction occurred owing to the fact that for the same period of time the wave moved by one cell to the left and the electron, by one cell to the right. Consequently, the backward-wave tube

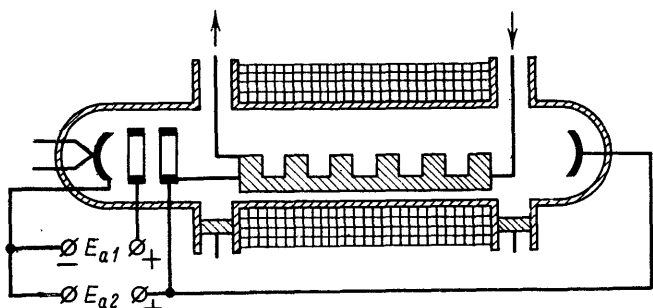


Fig. 10-49. Backward-wave amplifier

may operate effectively when the velocities of the electron stream and the electromagnetic wave are equal in magnitude and opposite in sign.

Now we may visualize a backward-wave amplifier (Fig. 10-49). Input signal should be applied to the end of the slow-wave structure next to the collector, and the amplified signal should be taken from its end next to the cathode. The uniform electron stream produced by the electron gun enters the intensive field of the amplified signal where it is modulated in velocity and then in density and gives up its energy to the backward wave.

It should be noted that backward-wave tubes are rarely used as amplifiers. Since electrons interact with the electric field only on separate parts of their path, the efficiency and amplification of these tubes are less than those of the travelling-wave tubes. Moreover, since the electron stream and the wave move in opposite directions a natural feedback exists in the device, thus presenting a serious problem of eliminating the oscillator self-excitation. Therefore, backward-wave tu-

bes are used primarily as self-oscillators. Such oscillators have only one high-frequency lead (Fig. 10-50).

The condition of effective interaction between the electron stream and the field in a backward-wave tube stated above may always be satisfied by the proper choice of accelerating voltage. This is essentially the self-excitation phase condition. The self-excitation amplitude condition reduces to the requirement for the tube gain that should exceed some cri-

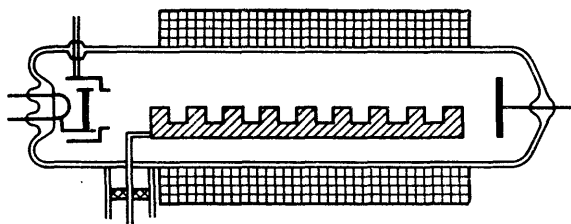


Fig. 10-50. Backward-wave self-oscillator

tical magnitude. The last requirement is met when a sufficiently great number of electrons are involved in the interaction process, i.e., the beam current in the tube should exceed some "starting" value.

Thus, if it is possible to change accelerating voltage no limitations exist in principle for the tuning range of a backward-wave tube. This is the chief advantage of this device. In practice, however, the tuning range is limited by a number of design and technological factors. The ratio of range extreme frequencies in some types of oscillators amounts to four or five.

The mechanism of initiating oscillations in backward-wave tubes is basically the same as in any other self-oscillator. After switching on supply voltages, a pulse of current charging distributed capacitances arises in the slow-wave structure. It produces oscillations of a continuous spectrum of frequencies so that damped waves start to propagate in different directions along the structure. That wave for which the self-excitation condition is fulfilled will modulate effectively the electron stream in velocity. The resultant electron bunches will deliver energy to this wave so that oscillations will rise. The system will gradually come into

a steady state since repulsive forces increase with the increasing density of electron bunches so that a further improvement of bunching ceases; in addition, losses in the system rise proportionally to the square of oscillation amplitude. Backward-wave tubes are used in the same ranges as travelling-wave tubes.

In the travelling-wave oscillators discussed above energy conversion occurs owing to the slowdown of electrons and the loss of their kinetic energy acquired earlier under the action of the accelerating field. Such oscillators are generally classified as *O-type oscillators*. A common feature of these oscillators is the difficulty of obtaining high efficiency since electrons cannot lose an appreciable fraction of their kinetic energy, otherwise the electron stream and wave velocities become unequal, i.e., the condition of effective interaction between the electric field and the wave will be disturbed. This is the principal disadvantage when designing high-power oscillators. To eliminate this drawback, another type of interaction between the electron stream and the field in crossed DC electric and magnetic field is used. The oscillators in which such an interaction is employed are called *M-type oscillators*.

10-8. M-TYPE TRAVELLING-WAVE OSCILLATORS

Let us consider the properties of electron motion in crossed magnetic and electric fields. For this purpose, imagine that we have two flat electrodes to which a potential difference is applied (Fig. 10-51). Assume that these electrodes are placed in a magnetic field so that its lines of force are directed away from us and perpendicular to the plane of the drawing. Suppose now that at some instant an electron appears at surface *K* which has a negative potential. Under the action of the electric field the electron will start moving towards surface *A* which has a positive potential. This motion, however, will be extremely complicated because of the magnetic field.

From elementary physics it is known that a DC magnetic field tends to displace a current-carrying conductor in the direction where the conductor field and the basic field have their lines of force in opposite directions (Fig. 10-52), i.e.,

in the direction where the resultant field becomes weaker. It is evident that force F acts on the electrons moving through the conductor rather than on the conductor itself (otherwise it will act on the conductor even when no current

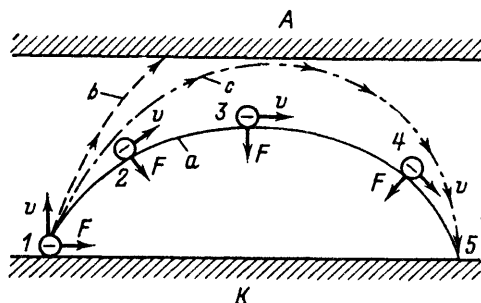


Fig. 10-51. Trajectories of electrons moving in crossed electric and magnetic fields

flows through it). The same process occurs when the electrons move in vacuum. At each point of the magnetic field the electrons will be influenced by the force which is perpendicular both to the direction of their motion and the magnetic field. The direction of this force can be found as follows. By using the corkscrew rule, the direction of the lines of force corresponding to the magnetic field produced by a moving electron is determined and then the direction along which the resultant field becomes weaker is found.

By this means, for instance, we may easily determine that the electron which moves from point 1 on surface K (Fig. 10-51) will deflect to the right. At some point 2 the electron velocity vector v is directed at a sharp angle to plane K and magnetic force F perpendicular to the velocity vector deflects the electron towards plane K. The same influence will be exerted upon the electron at points 3 and 4. As a consequence, the electron will return to the plane K (trajectory a). However, this will not occur at all times. If the electric

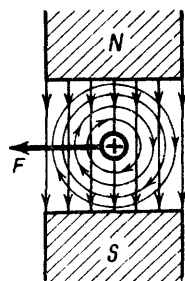


Fig. 10-52. Current-carrying wire in magnetic field

field is very intensive but the magnetic field is weak, trajectory bending will be negligible and the electron will fall on plane *A* (trajectory *b*). Finally, at some "critical" relationship between the electric and magnetic fields, the electron will pass in the immediate vicinity of plane *A* (trajectory *c*). Consequently, if plane *K* is an emitting electrode (cathode), the field relationship close to the critical one ($H > H_{crit}$) when the electrons pass near the surface of a

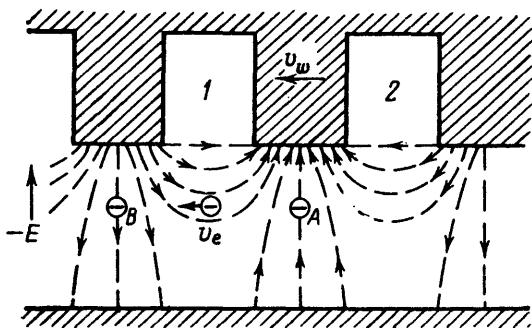


Fig. 10-53. Bunching of electron stream by travelling-wave field in M-type devices

positive potential (anode) may be easily established at all times.

Calculations show that under these conditions electron trajectories will coincide with the trajectories of the point on the rim of an imaginary wheel rolling over the cathode at the so-called "drift" velocity which is constant and defined as

$$v_e = \frac{E}{H} \quad (10-50)$$

where E is the electric-field intensity and H , the magnetic-field intensity. As some electrons fall on the cathode and the others move away from it, their vertical motion is mutually compensated and the average effect manifests itself as the motion of the electron cloud along the anode at a constant drift velocity.

Let us now assume that the anode and the cathode form a slow-wave system along which an electromagnetic wave (Fig. 10-53) is propagated, and that maximum retarding

and accelerating voltages* are applied at a given instant across slots 1 and 2, respectively. The electron stream moves at the same velocity and in the same direction as does the wave. While passing near the centre of slot 1, the electron is slowed down and gives up its energy to the wave. As this takes place, the electron starts to lag behind the travelling wave and enters the field whose configuration at the given instant corresponds, for instance, to point *A*. Here, the lines of force of the AC field bend (remember that they should be perpendicular to the surface of an ideal conductor, see Chapter V), assume a transverse direction, and then are summed up with those of the DC electric field. As a result, the electron velocity increases according to equation (10-50) and the electron is again introduced to the region of the maximum retarding field. If an electron appears at point *B*, its velocity will reduce since at this position the DC field is weakened by the wave field's transverse component in compliance with equation (10-50) so that this electron will also enter the region of the maximum retarding field. By reasoning in a similar manner, one may also show that electrons will flow away from under slot 2 where the accelerating field acts upon them, and move to the slot where they will be influenced by the retarding field or will fall on the cathode (this will be shown later).

Consequently, as distinct from O-type oscillators, the bunching of electron stream in M-type oscillators occurs under the action of a transverse, not longitudinal, component of the wave field. The wave field itself bunches the electrons and holds them in the region of the maximum retarding field where their energy is transferred effectively to the wave. Therefore, in contrast to O-type oscillators, here the velocity of the electron stream should precisely be equal to that of the wave in order that the energy exchange will occur in this region.

The same electrons cannot give up their energy over a long time period. Let us trace their path in order to understand the process as a whole. When an electron moves in

* For the sake of clarity, the arrows shown in Fig. 10-53 indicate the direction of forces acting on the electrons, not on the positive charges, therefore, the vectors of electric-field have, in this case, the opposite sign.

DC fields (see Fig. 10-51), its energy varies periodically. At the cathode the electron possesses the maximum potential energy with respect to the anode; at upper point 3 its energy decreases and is converted into the kinetic energy. The electron which has acquired a high kinetic energy may return to the cathode to point 5 where its potential energy will rise again to a maximum and the kinetic energy will become zero.

Let us now consider more complex phenomena when, aside from DC fields, the electrons are influenced by an AC field

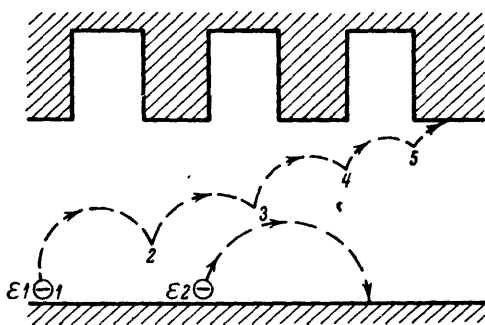


Fig. 10-54. Trajectories of electrons in M-type devices

of the wave travelling along the system. Assume that electron e_1 escaped from the cathode when it was in a wave retarding field (Fig. 10-54). Then, after losing some kinetic energy near the anode, the electron cannot return to the cathode and will come to a stop at point 2 on its way back to the cathode. If the electron drift velocity is equal to the wave velocity, the retarding field will at the same time arrive at the same point and, therefore, the electron will start moving once again to the anode and the cycle will repeat. In each cycle the electron while approaching the anode will lose some of its potential energy. Simultaneously, wave energy will increase by the same amount. On its last turn the electron strikes the anode, thus spending the remainder of its energy on heating the anode. The greater the number of turns made by the electron before it struck the anode, the less the remaining energy (the less the energy left in the electron on its last turn).

When describing the motion of the electron in DC fields, it was shown that the more intensive the magnetic field, the greater the force bending the electron trajectory, and, consequently, the smaller the turn radius and the greater the number of turns occurring between the anode and cathode. Hence, it is clear that the more intensive the magnetic field, the greater the efficiency of M-type oscillators. An increase in magnetic-field intensity, however, presents appreciable technical problems not only by itself but also because of the necessity of increasing proportionally anode voltage in order to maintain synchronism of the motion between the electron stream and the wave in compliance with equation (10-50).

The situation is quite different with electron ϵ_2 which escaped from the cathode when it was in an accelerating field (see Fig. 10-54). At the upper portion of its turn this electron acquires the kinetic energy which exceeds the potential energy lost by the electron and, therefore, it returns to the cathode and spends the remainder of its energy on heating the cathode.

Thus, the electrons giving up their energy to the wave in several cycles stay for a long time in the interaction space, whereas the electrons which take energy from the wave do it only once and return to the cathode after one convolution. Owing to this, the electron stream as a whole contributes its energy to the wave.

From the above stems the principal feature of M-type oscillators as in these particular oscillators the potential energy of electron stream is converted into wave energy. In these oscillators the expenditure of energy of electrons does not change the average stream velocity or disturb the conditions of synchronism. Consequently, there are no principal limitations on electronic efficiency of the oscillators of this type. In fact, efficiencies exceeding 80 per cent, are attainable in some M-type oscillators.

In most practical cases slow-wave structures of M-type oscillators are formed into a ring. Such a structure does not change the general pattern discussed above. In a cylindrical structure with DC electric and magnetic fields applied electrons will follow the trace of a point on a circle rolling over the cathode. The more intensive the magnetic field,

the less the radius of the circle. As in the flat (planar) structure, the trajectories of the electrons giving up energy to the wave will have the form of several turns (convolutions) terminating in the anode (Fig. 10-55). At the same time the electrons which tend to absorb energy will be quickly returned to the cathode (after the first turn).

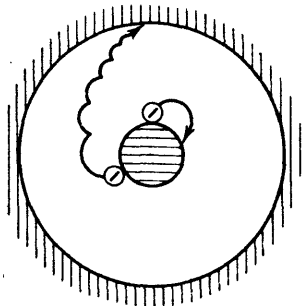


Fig. 10-55. Trajectories of electrons in M-type oscillators with circular structure

We now consider how these general ideas are embodied technically in actual devices. Fig. 10-56a illustrates the construction of a high power travelling-wave tube used as an M-type amplifier. The device having a cylindrical construction is placed in a magnetic field perpendicular to the plane of the drawing. The function of an electron gun consisting of cathode 1 and accelerating electrode 6 is to accelerate electrons which are then deflected by the magnetic field and follow along the slow-wave structure of the double-

comb type (also called the interdigital structure) 5 which serves as an anode. Direct voltage is applied between the anode and cold cathode 7. The input signal is fed through coaxial lead 4 located near the electron gun. The amplified

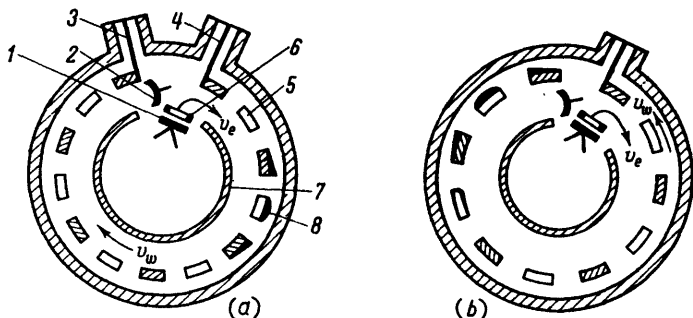


Fig. 10-56. High-power M-type devices

(a) travelling-wave amplifier; (b) backward-wave self-oscillator

comb type (also called the interdigital structure) 5 which serves as an anode. Direct voltage is applied between the anode and cold cathode 7. The input signal is fed through coaxial lead 4 located near the electron gun. The amplified

signal is taken at the end of the slow-wave structure from collector 3. Absorber 8 is arranged at the centre of the slow-wave structure.

Figure 10-56*b* illustrates the construction of a high-power backward-wave tube employed as an M-type oscillator. The construction is similar to that described above. It has the same slow-wave structure and only one RF lead located near the electron gun. Fluctuations in the electron stream initiate a number of waves travelling along the slow-wave structure, but only for one of them the condition of equal velocities and the opposite direction with respect to the electron stream is fulfilled. This wave is amplified and self-excitation occurs in the device. At centimetre waves oscillators of this type provide useful power up to 10 kW with efficiency exceeding 40 per cent and an electronic tuning range of 1 : 2.

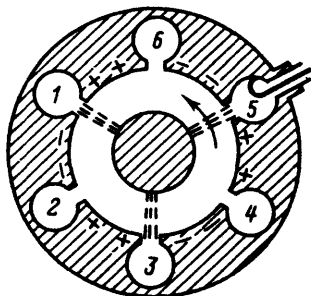


Fig. 10-57. Magnetron oscillator

One of the most widely used M-type travelling-wave oscillators is the *magnetron*. Its slow-wave structure is a chain of cavity resonators (Fig. 10-57). As distinct from the above oscillators, it does not have the electron gun (this is replaced by a massive cylindrical cathode) and the collector is also absent. The electron stream and the oscillatory system are of a closed type. Just as in the M-type oscillators discussed above, electron bunches ("spokes of a wheel") are formed in the electron cloud rotating about the cathode in the retarding electric field. A cut-away view of a magnetron is shown in Fig. 10-58.

The closed structure of the device gives rise to a new feature in its operation. A steady-state mode can be provided in the oscillator only if the wave running, for instance, from cavity resonator 1 (see Fig. 10-57) returns (after a complete turn around the slow-wave structure) to it in the same phase or in the phase differing from it by an integral number of 2π rad. While passing the cavity resonators, the current wave produces a voltage across each resonator.

This is shifted with respect to the current by some phase angle φ_n depending upon the resonator tuning relative to the frequency of the oscillations being generated. From the

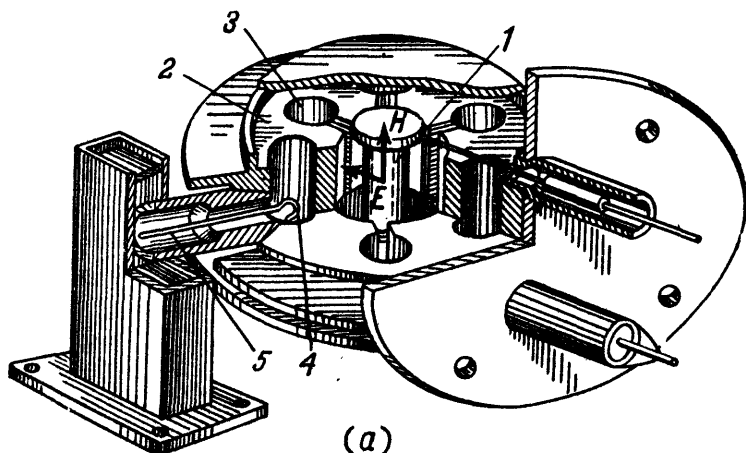
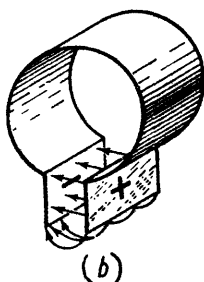


Fig. 10-58. Construction of magnetron (a) and its resonator (b)

1—cathode; 2—anode assembly; 3—resonator; 4—coupling loop; 5—HF lead



above phase condition it follows that the sum of these phase shifts should be

$$N\varphi_n = 2\pi n \quad (10-51)$$

where N is the number of cavity resonators; n , any integral number ($n=0, 1, 2, 3, \dots$), hence, the phase shift per each cavity resonator is

$$\varphi_n = 2\pi \frac{n}{N} \quad (10-52)$$

It is easy to verify that phase shift φ_n may assume different values only when $n \leq N/2$. For instance, if $n = N/2 + 1$, then

$$\varphi_{N/2+1} = 2\pi \left(\frac{1}{2} + \frac{1}{N} \right) = \frac{2\pi}{N} + \pi \quad (10-53)$$

which agrees except for the direction of reading with

$$\varphi_{N/2-1} = 2\pi \left(\frac{1}{2} - \frac{1}{N} \right) = \pi - \frac{2\pi}{N} \quad (10-54)$$

In fact,

$$\varphi_{N/2-1} - 2\pi = - \left(\frac{2\pi}{N} + \pi \right) = - \varphi_{N/2+1}$$

Consequently, as distinct from the above M-type oscillators having an open (straight-line) slow-wave structure, only $N/2+1$ different modes corresponding to $n=0, 1, \dots, N/2$ can be excited in the magnetron.

A different phase shift may arise in cavity resonators at a different frequency of the travelling wave which passes them. This means that different frequencies are excited in the magnetron at different modes. How may the desired mode be initiated in the magnetron? In Chapter V we have familiarized ourselves with the travelling-wave equation (5-4) which may be rewritten in the form

$$u_x = V \cos \left(\omega t - \frac{\omega x}{v} \right) \quad (10-55)$$

The second term in the brackets indicates how much the wave phase will change at distance x .

If the distance between two adjacent resonators is L and the phase shift between oscillations in them is φ_n (which should be considered as a result of travelling-wave delay), then from equation (10-55) we obtain

$$\varphi_n = \frac{\omega L}{v_n} \quad (10-56)$$

hence, taking into account equation (10-52), we get

$$v_n = \frac{\omega L}{\varphi_n} = \frac{\omega L N}{2\pi n} = \frac{L N f}{n} \quad (10-57)$$

The last equation indicates that the velocity of waves travelling along the slow-wave structure depends upon the mode being excited (upon number n).

Effective interaction between the wave and the electron stream can be obtained only if they move in synchronism, i.e., at the same velocity. Average velocity of the electron stream is determined by the relationship between the intensities of the electric and magnetic fields. Conse-

quently, in order to excite the desired mode, it is necessary to set up such an anode voltage at which the condition of synchronism between the electron stream and the wave will be satisfied, i.e.,

$$v_e = \frac{E}{H} = \frac{LNf}{n} \quad (10-58)$$

From this equation it follows that the velocity of electrons equal to that of the travelling wave decreases with the increasing number n . The lowest electron velocity is obtained at the maximum value $n=N/2$. In this case, oscillations in adjacent resonators will be shifted relative to one another by $\phi_{N/2}=\pi$, i.e., they will be 180° out of phase as can be seen from equation (10-52). These operating conditions are called π -mode. According to equation (10-58), a minimum anode voltage will be required to excite this mode. This is by far technically useful.

A detailed theory indicates that the most effective interaction between the electron stream and the travelling wave occurs when the oscillations are 180° out of phase. Therefore, the π -mode is used almost in all magnetrons. As energy is carried in the magnetron by the travelling wave which passes in succession all the resonators, there is no need to take energy from each resonator. Energy is drawn out of one resonator by an inductive coupling loop connected to an external feeder.

The magnetron is essentially a diode and, therefore, the only way to change the amplitude of oscillations in it is to vary its anode voltage. This would inevitably lead, however, to a change in the synchronism and to a transfer from one mode to another (this phenomenon is called mode skipping) and, consequently, to sudden changes in frequency, which is undesirable. Therefore, magnetrons are generally used in pulsed operation by applying supply voltage to the anode in the form of pulses with their tops as flat as possible.

Magnetrons are employed in the decimetre wave ranges but most frequently in the ranges of centimetre and millimetre waves. They are capable of producing peak powers (at centimetre waves) of about dozens of megawatts with a very high efficiency reaching 60 to 70 per cent.

Oscillations in the magnetron may be also excited due to interaction with the backward wave when the anode voltage is lower than in the π -mode.

Another M-type device developed for amplification of decimetre and centimetre waves is based on the interaction between the electron stream and the backward wave. This device is referred to as a *platinotron* (Fig. 10-59). A retarding system of the platinotron presents a two-wire line loaded with vane-type resonators. It has a cylindrical cathode as in the magnetron. The platinotron is not provided with an absorber since the line is matched to the input and output so that reflected waves are absent. The input signal produces waves travelling in the slow-wave structure with phase velocity v_w . When the synchronism condition is satisfied (i.e., when the electron velocity v_e and wave velocity v_w are equal), the energy interaction discussed above occurs between the electron stream and the backward wave.

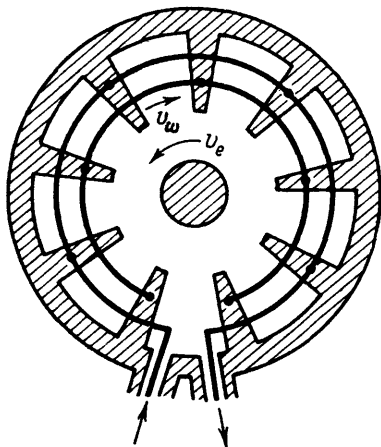


Fig. 10-59. Platinotron used as amplifier (amplitron)

When used as an amplifier, the device is referred to as an *amplitron*. It has a rather low gain (up to 20 times) but provides a very high efficiency (up to 80 per cent). Due to the closed electron stream, the bandwidth of the amplitron is less than that of the travelling-wave tube, though it still amounts to 10 per cent.

10-9. TRANSISTOR OSCILLATORS AND AMPLIFIERS

About 10 to 50 per cent of the power fed to electron valves is spent in the filament (heater) circuit and, therefore, investigations are conducted over a long period to find a way of replacing electron valves with devices having no filament.

Transistors proved to be such devices which made it possible to increase drastically the efficiency of oscillators and amplifiers. Their use allowed also considerable advances to be made in increasing the reliability of radio equipment since burning of filaments and loss of cathode emission are the major causes of valve failure.

Comparison between the characteristics of electron valves and transistors has revealed their identical nature (see Chap-

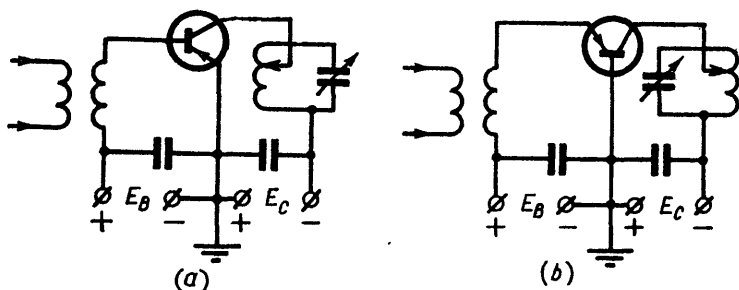


Fig. 10-60. Separately-excited oscillators employing transistors
(a) common-emitter oscillator; (b) common-base oscillator

ter VIII), which offers the basis for their identical application. In the latter, the collector functions as the anode, the base as the grid and the emitter as the cathode.

The major difference between analogous circuits which employ transistors having different types of conductance consists in the polarity of supply sources. If we deal with *P-N-P* transistors, collector and base currents may be discontinued by applying a small positive or zero bias to the base. Therefore, when transistors are operated at large cutoff angles ($\theta \approx 90^\circ$), positive bias voltages have to be frequently applied to the base whereas in class C operation, bias voltage should be negative, just like in valve oscillators. In both cases, a negative voltage with respect to the emitter is applied to the collector of such transistors. Fig. 10-60 illustrates the schematic diagrams of common-emitter and common-base amplifiers similar to the common-cathode and common-grid valve amplifiers, respectively.

Owing to high transconductance which is, on average, 200 mA/V (i.e., by an order of magnitude more than in

valves) and low base currents, a common-emitter transistor amplifier provides much greater gain than a common-cathode valve amplifier. Amplification, however, of both the common-base oscillator and common-grid oscillator is not great.

Among the disadvantages of transistor amplifiers are, first, their strong temperature dependence which causes the designers to complicate circuits with special temperature-compensating devices, and, secondly, deterioration of their amplifying properties with the increasing frequency.

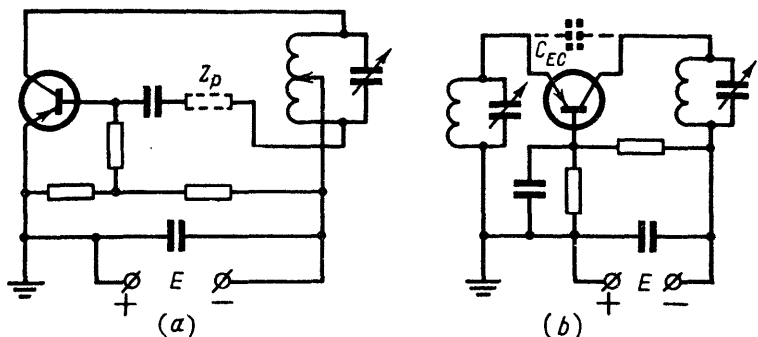


Fig. 10-61. Transistor self-oscillators

(a) tapped-coil oscillator; (b) two-circuit self-oscillator employing common-base transistor

The circuits of transistor self-oscillators are also analogous to those of valve self-oscillators (Fig. 10-61). At relatively high frequencies, however, it is often necessary, due to a finite mobility of charge carriers, to introduce an additional phase-shifting network $R_p C_p$ or a single reactance Z_p into the feedback circuit to fulfil the phase condition of self-excitation.

Combined-bias circuits are used, as a rule, in transistor self-oscillators in order to obtain soft self-excitation and small cutoffs under steady-state operating conditions. An initial bias, which makes it possible to obtain a sufficiently high transconductance at the instant of switching on is produced by applying some voltage from a collector supply source through a voltage divider. Under steady-state conditions the base current flowing through an automatic-bias network shifts the oscillator in class C operation.

10-10. NEGATIVE-RESISTANCE OSCILLATORS

In a single tuned circuit sustained oscillations may be produced only if losses in it are zero, i.e., its resistance is equal to zero. As any actual tuned circuit offers some resistance, the generation of sustained oscillations in the tuned circuit of a valve oscillator may be considered as a result of introducing some "negative resistance" into the circuit, which neutralizes its positive resistance. The valve whose grid is fed with voltage in phase with the voltage across the tuned circuit functions in the valve oscillator as such a

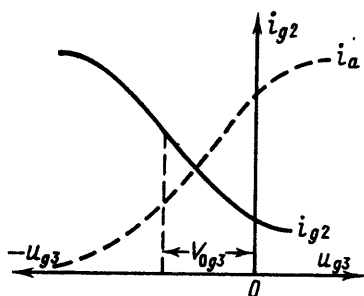


Fig. 10-62. Screen-grid current characteristic curve as a function of suppressor voltage

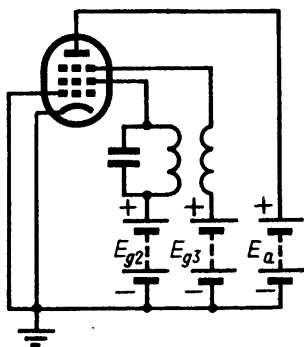


Fig. 10-63. Transiton oscillator

negative resistance. If phase relationships in the oscillator are maintained in compliance with the graphs shown in Fig. 10-3, the voltage between the anode and the cathode decreases with the increasing anode current (as distinct from an increase in voltage across an ordinary resistor with the increasing current). This is a distinctive feature of a negative-resistance element. The valve itself offers, however, positive resistance. It displays negative resistance properties only with respect to the first harmonic of anode current, to which the circuit is tuned, when an exciting voltage is fed to it from another oscillator or from the valve anode circuit through a feedback path.

There are other ways of obtaining the negative resistance in electron valves. For instance, a decrease in screen-grid

current is generally observed in multigrid valves with the increasing suppressor-grid voltage inspite of a simultaneous increase in screen-grid voltage due to redistribution of electron flow between the anode and the screen grid.

In Fig. 10-62 screen-grid current as a function of suppressor-grid voltage is shown for a typical multigrid valve. If we connect a tuned circuit between the screen grid and the cathode of such a valve and apply a part of voltage from the circuit to the suppressor grid through a feedback path so that suppressor-grid potential will rise with the rising potential at the screen grid, then the screen-grid current will decrease and, consequently, the cathode-screen grid portion of the valve will function as a negative resistance. This phenomenon underlies the operation of the so-called *transitron oscillators* (Fig. 10-63).

Some devices by their nature exhibit negative resistance for alternating current. For instance, an arc discharge has a drooping volt-ampere characteristic typical of negative-resistance elements. Arc temperature and the number of ions in an arc gap increase with the increasing current. As a consequence, the resistance of the arc gap and the potential difference between its electrodes decrease accordingly. This property of the arc discharge was used in high-frequency arc oscillators before the advent of valve oscillators. In these oscillators arc discharges occurred in the circuit where high-frequency oscillations were sustained. Oscillators are frequently referred to as negative-resistance devices in those cases when, as in the arc oscillator, no feedback is required for imparting the negative-resistance properties to the element which maintains oscillations in the circuit.

10-11. SINEWAVE RC OSCILLATORS

Sinewave oscillations are produced in the circuits containing reactive elements of two types, namely, inductances and capacitances (LC circuits). As shown in Chapter II, any disturbance of electrical equilibrium of RC or RL circuits results in the flow of nonsinusoidal currents. This, however, does not mean that sinewave oscillations cannot be separated in the self-oscillating system employing such circuits. It will be recalled that nonsinusoidal currents may

be presented in the form of a sum of an infinite series of harmonic (sinewave) components having different frequencies. To excite oscillations of only one frequency in an oscillatory system, it is necessary to build the system so that the conditions of self-excitation will be satisfied only for one harmonic. It is simpler to attain this by composing the feedback circuit so that the phase condition of self-excitation will be satisfied for only one frequency.

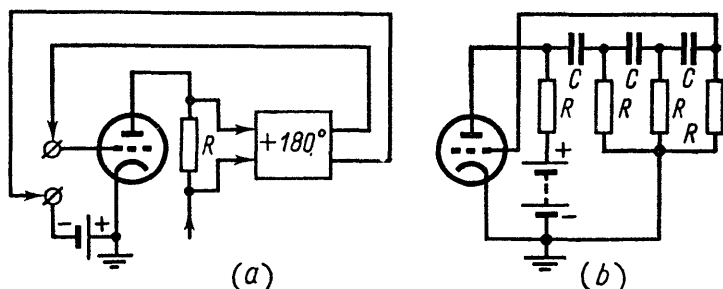


Fig. 10-64. RC self-oscillator

(a) self-oscillator obtained by using feedback in resistance-coupled amplifier; (b) schematic diagram of self-oscillator

A resistance-coupled valve amplifier may be converted into the self-oscillator if the amplified voltage is fed partially from its anode load to the control grid 180° out of phase with the anode AC voltage (Fig. 10-64a). RC networks may be used for obtaining the desired phase shift of the voltage fed from the load resistor to the control grid. Each such network may provide, in principle, a phase shift not exceeding 90° though, in practice, a phase shift of about 60° is attainable. Therefore, a network consisting of three RC sections is generally connected into the feedback circuit in order to obtain a phase shift of 180° (Fig. 10-64b). The voltage taken from the last RC section is applied to the control grid of the valve.

The phase shift produced by each RC section and the RC network as a whole depends upon its parameters and the operating frequency. At other frequencies, therefore, the desired phase shift will differ from 180° so that oscillations will not be excited. In the above oscillatory system the

frequency of the oscillations being generated will be

$$\omega = \frac{1}{\sqrt{6}RC} \quad (10-59)$$

In an oscillator of this type as compared with a tuned-circuit oscillator the operating frequency can be varied over a wider range since in the latter it is inversely proportional to the square root of capacitance, provided the same variable capacitor is used in both oscillators. This is the great advantage of the RC oscillators. If, for instance, the variable capacitor has the ratio $C_{\max}/C_{\min}=4$, then the frequency can be varied four times in the RC oscillator and only two times in the LC oscillator.

Among other advantages of RC oscillators are also their small weight, size and cost, which are particularly noticeable at low frequencies where iron-cored chokes are extremely bulky and expensive. A great variety of sinewave RC oscillators are developed and used at the present time. RL oscillators may be built in a similar way but they have large weight, size and lower frequency stability. For these reasons, they have found no practical use in radio equipment.

10-12. FREQUENCY PULLING IN SELF-OSCILLATORS

Early valve radio-transmitters employed basically a single-stage circuit which consisted of a self-oscillator connected directly with an antenna. In order to transmit the maximum power to the antenna, the coupling between the oscillator and the antenna was selected as tight as possible. It was found when tuning the antenna circuit that oscillations ceased or oscillator tuning became ambiguous, i.e., the frequency being generated depended upon the tuning procedure used.

Let us analyze the cause of this phenomenon. If we replace the antenna with an equivalent circuit, it will be easily seen from the oscillator circuit represented in Fig. 10-65 that its load consists of two coupled circuits. Assume that the circuit connected between the anode and cathode of the valve is tuned by capacitor C_1 to a fixed frequency f_1 , and

that the external-circuit frequency, f_2 , is varied with the aid of capacitor C_2 .

The frequency of the oscillations being generated coincides with the resonance frequency of the oscillatory system whose natural frequencies are determined by the coupling frequencies of the two circuits. The coupling frequencies vary depending upon the value of coupling, X_m , and the difference between them (see Fig. 3-8).

At the instant the oscillator is switched on, weak oscillations of both frequencies are excited in its oscillatory system.

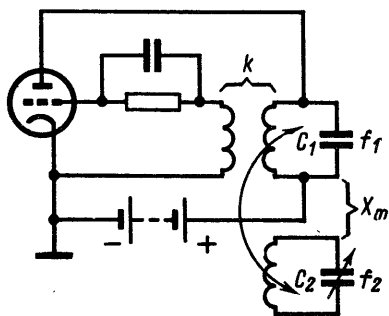


Fig. 10-65. Self-oscillator loaded with two coupled circuits

The presence of oscillations of one frequency in the oscillator, however, impedes excitation of other frequencies in it. This is explained by the fact that all self-oscillators operate in a non-linear mode, i.e., in the region with a varying transconductance. Therefore, the average transconductance of the valve during oscillations proves always to be less than its static transconductance. It is

known from general conditions of self-excitation (see equation 10-32) that the less the valve average transconductance, the greater has to be the feedback ratio at which a new frequency may be excited. Furthermore, the presence of oscillations of one frequency leads to the appearance of a retarding field in the valve not at those instants when electrons associated with the pulses following at another frequency can give up most effectively their energy to the circuit. Therefore, the rapidly rising oscillations of the frequency at which losses in the system are less, impede excitation of oscillations of the other frequency and the latter attenuate quickly.

Let us consider phenomena which occur in the oscillator when its external circuit is tuned. This results in the change of coupling frequencies one of which lies below and the other above the natural frequencies of both tuned circuits. If the natural frequency of the external circuit is lower than that

of the anode circuit, the amount of energy being transferred from the anode circuit to the external circuit will be greater at the lower coupling frequency ω_l than at the upper one ω_{up} , since the difference $\omega_l - \omega_2$ is less than $\omega_{up} - \omega_2$ (see Fig. 3-8). Therefore, oscillations are excited at the upper coupling frequency at which losses in the anode circuit are less.

If the external circuit is tuned to the frequency which is higher than that of the anode circuit, it will be detuned with respect to the upper coupling frequency, $\Delta\omega = \omega_{up} - \omega_2$, less than with respect to the lower coupling frequency, $\Delta\omega = \omega_l - \omega_2$. Therefore, anode-circuit losses associated with energy transfer to the external circuit will become greater at the upper coupling frequency so that oscillations at the lower coupling frequency will be excited in the oscillator. As seen from Fig. 10-66a, when $\omega_2 < \omega_l$ an increase of the frequency ω_2 results in a decrease of external-circuit detuning relative to the operating frequency and an increase of current I_2 through the external circuit. When $\omega_2 > \omega_l$, an increase of the frequency ω_2 leads to an increase of external circuit detuning with respect to the lower coupling frequency which is excited in the oscillator and a decrease of current in the external circuit.

If coupling between the anode and external circuits is tight enough, energy transfer to the external circuit is especially intensive when the latter is detuned little with respect to the frequency being generated. As a consequence, the equivalent resistance of the anode circuit is decreased appreciably. Therefore, as seen from formula (10-30), a preset feedback ratio may turn to be insufficient for maintaining self-excitation, and oscillations may cease. The phenomena occurring in the oscillator are illustrated for this case by the curves shown in Fig. 10-66b.

It is evident that oscillation stopping may be eliminated by increasing the feedback. If the external-circuit frequency is increased, the processes occurring in the oscillator will proceed as follows. If $\omega_2 \leq \omega_l$, oscillations will be excited at the upper coupling frequency. Despite the fact that more favourable conditions (from the viewpoint of losses) exist in the circuit for exciting oscillations at the lower coupling frequency, they are not still excited when transferring thro-

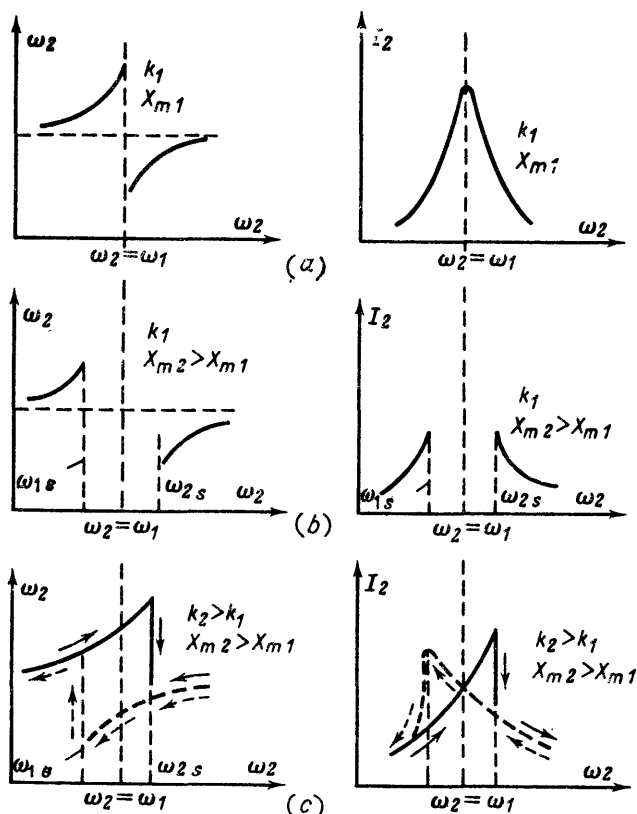


Fig. 10-66. Frequency pulling in two-circuit self-oscillator

ugh frequency ω_1 up to stopping frequency ω_{2s} since the presence of upper-frequency oscillations reduces the average transconductance, thus preventing their appearance. Only at frequency $\omega_2 = \omega_{2s}$, much more favourable conditions are produced in the circuit for exciting the oscillations at the lower coupling frequency, which results in a sudden change (skip) of oscillation frequency (Fig. 10-66c). Before the sudden change in frequency an increasing energy transfer to the external circuit causes an increase in current flowing through it. Naturally, when the frequency of the external

circuit is decreased, *pulling* of oscillator frequency at the lower coupling frequency (dashed lines in Fig. 10-66c) and then the skip to the upper coupling frequency are observed.

Thus, the external circuit tuning range, ω_{1s} to ω_{2s} , is formed within which the operating frequency of the oscillator depends upon how it was tuned. This undesired phenomenon may be encountered in single-stage transmitters which are used in a variety of pulse radio devices. In multistage transmitters the oscillator is always coupled to the succeeding stages and, therefore, any attempt to increase coupling between them above some limit may lead to the pulling of oscillator frequency. In some devices the pulling effect is used for frequency stabilization.

It might be well to point out that the pulling effect may occur only in those oscillators where one of the coupled circuits functions as an "external" circuit from which energy is not transferred back to the self-oscillator circuit. In two-circuit common-anode, common-cathode or common-grid oscillators the second tuned circuit is incorporated in the feedback circuit. As can be seen from the curves shown in Fig. 10-29, in these oscillators the signs of feedback ratios are opposite at the two coupling frequencies and, therefore, only one of them may be excited. In such oscillators, the pulling effect may be observed only in those cases when the tuned circuit determining the frequency of the oscillations being generated is coupled tightly to the external circuit.

10-13. SELF-OSCILLATOR LOCK-IN

If a self-oscillator generating oscillations at frequency ω_{osc} is exposed to the action of a sufficiently high external (injection) EMF of frequency ω_{inj} differing slightly from ω_{osc} the oscillator frequency may change and become equal to that of the external EMF after some transient period. This phenomenon is known as a *full lock-in* of the oscillator. In some instances frequency lock-in is also referred to as *injection synchronization*. At a given level of injection signal the lock-in is observed within a frequency range lying around the natural frequency of the oscillator. This frequency band is called the *lock-in range*. At a somewhat larger difference between the two frequencies a partial change in oscillator

frequency is observed. This phenomenon is known as *partial lock-in* of the oscillator. As the difference is increased further, the injection signal does not change the oscillator frequency any more.

In Fig. 10-67 the difference between the oscillator operating frequency and the injection-signal frequency is shown as a function of the latter. In the lock-in range this difference is zero, in the region of the partial lock-in it is different from zero

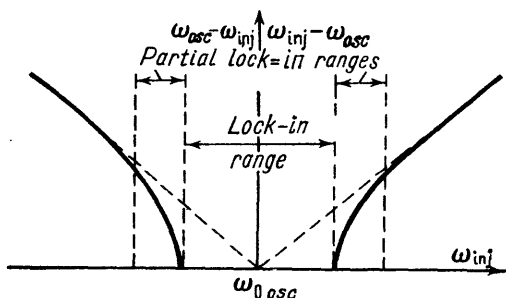


Fig. 10-67. Self-oscillator lock-in

and outside the partial lock-in region the difference rises linearly since the frequency $\omega_{0 osc}$ ceases changing. This may be explained as follows. Because of feedback inherent in any self-oscillator an external signal will always appear at the control grid of the valve no matter whether it acts on one or another of the self-oscillator circuits. If the frequency of the external signal does not differ essentially from that of the self-oscillator and the amplitude of oscillations induced by it is comparable to that of the voltage across the grid, the *beating* of these two frequencies arises in the grid circuit. At some instant the peak positive amplitudes of both signals coincide (Fig. 10-68), they are summed up, and the amplitude of the resultant oscillation is equal to the sum of their amplitudes. After one cycle one of the signals will again reach its maximum, whereas the amplitude of the second signal at the same instant will be less and, therefore, the resultant amplitude will be less than in the first case.

Phase shift between the two signals will increase from one cycle to another and the amplitude of the resultant oscilla-

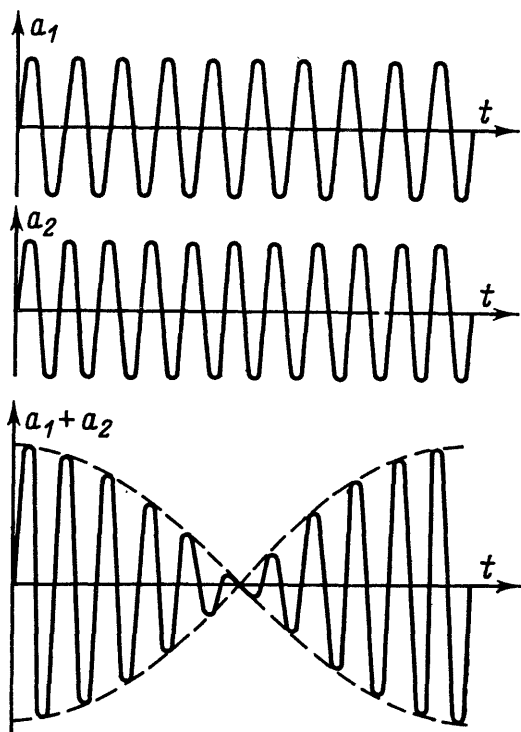


Fig. 10-68. Beats resulting when two oscillations of different frequencies are summed up; frequency difference is small

tion will decrease. When the phase shift exceeds 120° , the signals will weaken each other and at a phase shift of 180° the resultant amplitude will be equal to the difference between the peak amplitudes of the two signals. A further increase in the phase shift will lead to a gradual increase of the resultant oscillation. When the phase shift reaches 360° , the signals will sum up again. Thus, a new oscillation with a periodically increasing or decreasing amplitude arises as a result of summation of two signals having different frequencies.

Let us describe the above phenomenon in mathematical terms in order to determine the frequency and period of the

resultant oscillation. If two high-frequency signals of the same amplitude, $a_1(t) = A \sin 2\pi f_1 t$ and $a_2(t) = A \sin 2\pi f_2 t$, are summed up in the circuit, then the result will be

$$a_1(t) + a_2(t) = A (\sin 2\pi f_1 t + \sin 2\pi f_2 t)$$

By using a well-known trigonometrical formula for the sum of two sines, namely,

$$\sin \alpha + \sin \beta = 2 \cos \frac{\alpha - \beta}{2} \sin \frac{\alpha + \beta}{2}$$

we obtain

$$a_1(t) + a_2(t) = 2A \cos 2\pi \frac{f_1 - f_2}{2} t \times \sin 2\pi \frac{f_1 + f_2}{2} t \quad (10-60)$$

If frequency f_1 is close to frequency f_2 , the frequency of the first factor is low, and it may be considered as the instantaneous amplitude of an oscillation whose frequency will be

$$f = \frac{f_1 + f_2}{2}$$

Thus, as a result of beating of two high-frequency signals, a new oscillation arises whose frequency is equal to the average of their frequencies, i.e.,

$$f = f_{av} = \frac{f_1 + f_2}{2} \quad (10-61)$$

This new oscillation has a varying amplitude whose peak value is repeated at a frequency equal to the difference of the two original frequencies, i.e.,

$$F = f_1 - f_2 \quad (10-62)$$

The equation of the resultant oscillation indicates that when its amplitude passes through zero, its phase is reversed since $\cos \frac{f_1 - f_2}{2} t$ changes its sign.

When the amplitudes of the two signals, A_1 and A_2 , are not identical, the amplitude of the resultant oscillation varies from a maximum value, $A_{\max} = A_1 + A_2$, when the two original signals are in phase to a minimum value, $A_{\min} = A_1 - A_2$, when their phase difference amounts to π . In this case, the period of amplitude variations of the resultant oscillation can be determined from the following conside-

rations. If the original signals are in phase at an initial instant, they will be again in phase after time T during which one of them will perform n , and the second one $n+1$ cycles, i.e., $T=nT_1$ and $T=(n+1)T_2$. Equating the right-hand sides of these equations, we obtain

$$n = \frac{T_2}{T_1 - T_2}$$

Substitution of this expression into the first equation gives the period of amplitude variations of the resultant oscillation which is

$$T = nT_1 = \frac{T_2}{T_1 - T_2} T_1 = \frac{1}{\frac{1}{T_1} - \frac{1}{T_2}}$$

hence, the frequency of amplitude variations will be

$$F = \frac{1}{T} = f_1 - f_2$$

i.e., it is equal to the difference in the two original frequencies.

Let us return to the phenomenon of frequency lock-in in the self-oscillator. From the above considerations it follows that the self-oscillator signal and the injection signal will produce a new oscillation whose amplitude will vary at the rate $\omega_{0osc} - \omega_{inj}$ and its frequency will be equal to

$$\omega'_{rsl} = \frac{\omega_{0osc} + \omega_{inj}}{2}$$

As frequencies ω_{0osc} and ω_{inj} are close to each other, their difference is much less than the frequency of the resultant oscillation. Therefore, we may neglect changes in its amplitude by considering small time intervals (of the order of several oscillation cycles).

Anode current is determined basically by the voltage across the control grid and, therefore, it will also change with the frequency ω_{rsl} . While flowing through the anode tuned circuit, the anode current will set up a voltage of a new frequency ω'_{rsl} across it. It can be easily seen that if $\omega_{inj} > \omega_{0osc}$, the new frequency equal to the average of ω_{inj} and ω_{0osc} will be higher than ω_{0osc} but less than ω_{inj} ,

i.e., the frequency being generated will rise. Oscillation of the new frequency ω'_{rsl} will produce beats with the injection signal, thus resulting in the oscillation generated at frequency

$$\omega''_{rsl} = \frac{\omega'_{rsl} + \omega_{inj}}{2}$$

which will be higher than ω'_{rsl} but lower than ω_{inj} (Fig. 10-69) so that the oscillator frequency will rise still more.

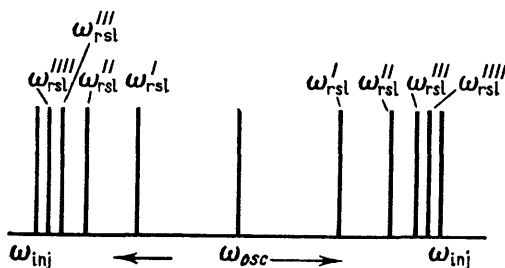


Fig. 10-69. Variations in self-oscillator frequency during lock-in process

The process of the frequency rise will develop quite rapidly until the oscillator frequency coincides almost accurately with that of the injection-signal $\omega_{rsl}^{(n)} = \omega_{inj}$ after some n th cycle, and no further variations of oscillator frequency will occur. The same happens when the frequency of the injection signal is less than that of the oscillator but in this case the oscillator frequency will decrease until it coincides with the injection signal frequency. The graph in Fig. 10-69 represents schematically the results that could be obtained if one measured the instantaneous magnitude of frequency at separate moments of time. In actual self-oscillators, however, frequency changes continuously.

The greater the difference between the frequency of the injection signal and the natural frequency of the self-oscillator, the greater the detuning of the anode circuit with respect to the injected frequency. Beginning with some detuning value, conditions for the excitation of frequency ω_{inj} are not satisfied and the oscillator returns to free-running mode. This does not mean, however, that the influence of the in-

jection signal discontinues completely. The above process of the gradual approach of the frequency being generated to that of the injection signal will stop in a certain stage and some intermediate frequency will be set up in the circuit. This frequency can be determined from the following general formula:

$$\omega_{rsl}^k = \frac{\omega_{rsl}^{(k-1)} + \omega_{inj}}{2}$$

The conditions of self-excitation for this intermediate frequency will be more favourable than for the next beat frequency. The former will be the closer to the injection-signal source frequency, the less the source is detuned with respect to the oscillator original frequency.

The above phenomenon is referred to as a *partial lock-in*. It is observed within some detuning region beyond which the partial lock-in ceases as shown in Fig. 10-69.

The lock-in effect is used in a variety of radio devices, for instance, to synchronize an oscillator from another more stable oscillator, to separate and amplify the desired signal when many signals of different frequencies are present in the circuit, to multiply frequency by synchronizing the oscillator with higher harmonics of the exciting voltage and for many other purposes.

10-14. NONSINEWAVE OSCILLATORS

Nonsinewave oscillators are used extensively in modern radio equipment for transmitting intelligence in the form of coded pulse trains, for simultaneous triggering of different devices, for successive actuating of different circuits at certain time intervals and for other purposes. The electrical signals of different waveforms other than a sinewave are called *relaxation oscillations*.

A schematic diagram of a relaxation oscillator is shown in Fig. 10-70a. A neon lamp, *NL*, with a shunt capacitor *C* is connected to a DC source *E* through a resistor *R*. When the circuit is turned on, the capacitor charges from the supply source through resistor *R*. As soon as voltage across the capacitor reaches the starting potential of the neon lamp, an ion discharge occurs in the lamp, its internal resistance becomes

very low (usually of the order of hundreds of ohms) and the capacitor starts to discharge through the neon lamp. The capacitor discharges until its voltage drops down to the extinction potential which is below the starting potential. The resistor R serves to limit the current flowing from the supply source and prevents the discharge from being maintained in the lamp. Therefore, the neon lamp goes out and the capacitor starts to charge once again.

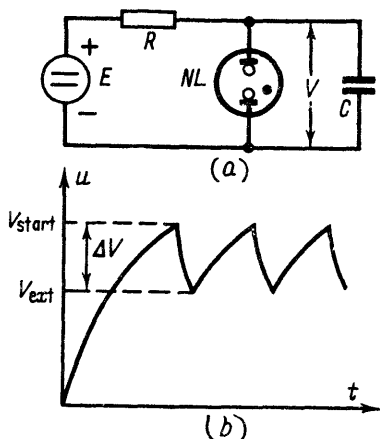


Fig. 10-70. Relaxation oscillator employing a neon lamp

The capacitor charges and discharges through the resistor by an exponential law (see Chapter II). When the capacitor is charged through a high resistance of the resistor R , voltage rises slowly, whereas when the capacitor is discharged through a small internal resistance of the conductive lamp, the process occurs many more times quicker. This process results in sawtooth variations of voltage

across the capacitor as illustrated in Fig. 10-70b.

Despite its simplicity, such an oscillator is used comparatively rarely due to instability of neon-lamp starting and extinction potentials which depend upon temperature, ambient illumination, a previous state of the lamp and upon other accidental reasons as well as due to a low amplitude of oscillations being generated. This amplitude is defined as

$$\Delta V = V_{start} - V_{ext}$$

where V_{start} is the starting voltage and V_{ext} , the extinction voltage of the neon lamp. More stable nonsinewave oscillators may be built using electron valves or semiconductor devices.

Two triodes are employed, for instance, in the circuit of a relaxation oscillator known as a *multivibrator* (Fig. 10-71).

The excitation of oscillations in this circuit may be described as follows.

Assume that at some instant of time the multivibrator is in electrical equilibrium, i.e., that equal currents flow through its two arms. This state, however, cannot exist for a long period of time because of small accidental changes of electron flow in electron valves. Let us now assume that at some instant the anode current of valve V_1 slightly rises. Then, the voltage drop across resistor R_{a1} will increase. As this takes place, the sum of voltages across resistor R_{a1} and valve V_1 will remain constant at all times and equal to the supply-source voltage, i.e.,

$$V_{Ra1} + V_{a1} = E_a$$

where V_{Ra1} is the voltage drop across resistor R_{a1} , and V_{a1} , the anode voltage across valve V_1 .

Consequently, if V_{Ra1} has increased, V_{a1} has to decrease. As a result, capacitor C_1 connected in parallel with valve V_1 through resistor R_{g2} will start to discharge. Its discharge current i_{dis} passes from the bottom upwards through the resistor R_{g2} connected into the grid circuit of valve V_2 and increases the grid negative potential with respect to the cathode. This results in a decrease of anode current in valve V_2 . In its turn, the decrease of anode current in valve V_2 leads to a decrease in the voltage across resistor R_{a2} . Therefore, capacitor C_2 starts to charge. While passing from the top downwards through the resistor R_{g1} , its charging current i_{ch} increases the grid voltage of valve V_1 with respect to its cathode, which aids in a further rise of anode current in valve V_1 . It should be noted that the positive voltage across the grid of valve V_1 gives rise to a grid current. Therefore, capacitor C_2 will charge also through the grid-cathode portion of valve V_1 . From all this it follows that positive feedback exists in the multivibrator and this leads to a rapid increase of current in valve V_1 and a decrease of current in valve V_2 .

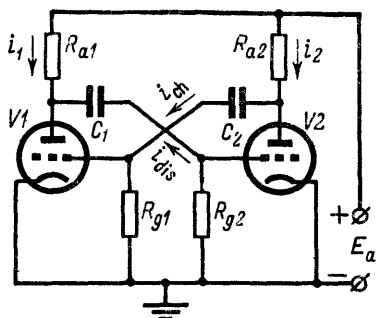


Fig. 10-71. Multivibrator

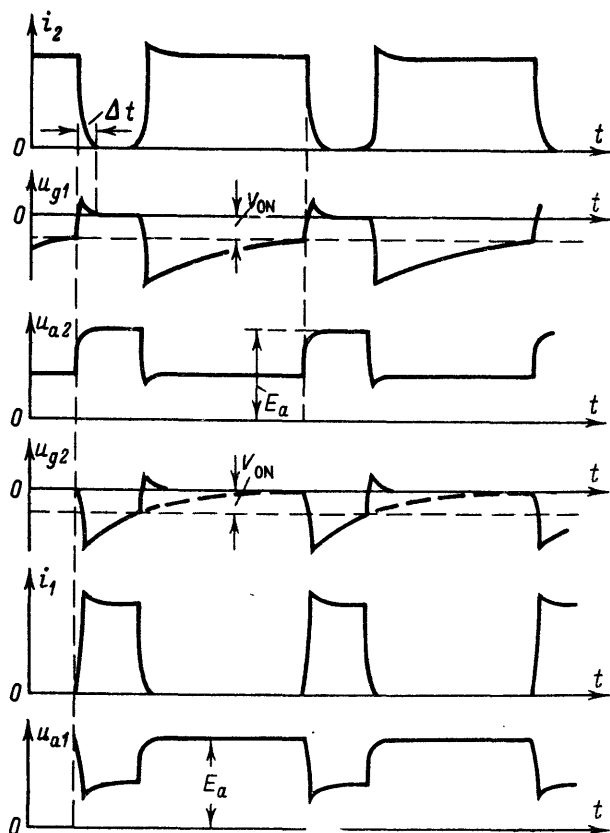


Fig. 10-72. Processes in multivibrator

The first stage of the process terminates in complete cutoff of valve V_2 and the passage of a large current through valve V_1 . The multivibrator, however, will not remain for a long time in this state. The capacitor C_1 will start to discharge through conductive valve V_1 and the current through the capacitor will decrease, thus resulting in a gradual rise of voltage across the grid of valve V_2 . At some instant the grid voltage will exceed the cutoff voltage of valve V_2 , and current will appear in its anode circuit. This will give rise

to a voltage drop across resistor R_{a2} and the anode voltage of valve V2 will start to decrease. As a consequence, capacitor C_2 will start to discharge through resistor R_{g1} , thus decreasing the voltage across the grid of valve V1, which in its turn lessens the anode current in this valve and increases its anode voltage. Simultaneously, capacitor C_1 starts to charge through resistor R_{g2} , thus increasing the voltage across the grid of valve V2. As a result, the current flowing through this valve will increase. Thus, an avalanche-type increase of current in valve V2 and a decrease of current in valve V1 take place. The circuit is, so to say, "flip-flopped", i.e., the processes which take place first in the valve V1 occur then in the valve V2 and vice versa. Further, the process described will be repeated and the multivibrator will produce a periodic train of relaxation oscillations.

Let us examine now the waveform of oscillations in different circuits of the multivibrator (Fig. 10-72). At the instant when valve V2 is cut off, capacitor C_2 will not yet stop to charge as the current jump occurs very quickly in the valve and the charging-circuit time constant determined by the value of capacitance C_2 and the values of resistances R_{a2} and R_{g1} is very large (it should be taken into account that the last resistor is shunted by a relatively low input impedance of the conductive valve V1).

The resistances of resistors R_{a1} and R_{a2} are selected to be not very high. Therefore, capacitor C_2 will be charged in a short time interval Δt after the cutoff of the valve begins. Voltage across the anode of valve V2 will reach in this interval its maximum value equal to source voltage E_a and the voltage across the grid of valve V1 will drop down to zero since charging current ceases so that the voltage drop across resistor R_{g1} disappears.

Simultaneously, capacitor C_1 starts to discharge through valve V1 and resistor R_{g2} since by this moment valve V1 has become fully conductive and the voltage drop across resistor R_{a1} has increased. Since resistance R_{g2} is selected to be very high, the discharge process progresses comparatively slowly. At the initial moment discharge current is a maximum and then decreases by an exponential law. Therefore, voltage across resistor R_{g2} , i.e., voltage u_{g2} applied to the grid circuit of valve V2 will change accordingly.

The current flowing through resistor R_{a1} rises quickly to a maximum and then decreases somewhat by an exponential law since grid voltage u_{g1} drops down to zero. The current remains at this level until the circuit changes its state. Voltage across the anode of valve V1 changes accordingly. Thus, we may write that

$$u_{a1} = E_a - i_1 R_{a1}$$

At the instant when the grid voltage of valve V2 rises up to voltage V_{ON} due to the decrease in discharge current of capacitor C_1 , valve V2 becomes conductive abruptly (is turned on), whereas valve V1, on the contrary, becomes cut off and all the processes in the circuit progress similarly to those described above but in the reverse direction.

At the instant valve V1 is cut off, the grid potential of valve V2 rises abruptly at the expense of the voltage produced by the charging current of capacitor C_1 across resistor R_{g2} and it then drops down to zero as the charging current ceases. A high negative voltage is produced by the discharge current of capacitor C_2 across resistor R_{g1} when valve V2 becomes conductive. Grid voltage increases as discharge current decreases. If the grid voltage exceeds the valve turn-on voltage, a new current jump occurs and all processes are repeated. Anode currents and voltages of valves V1 and V2 remain constant in the intervals between the jumps if we neglect short time intervals during which transient processes following immediately the circuit flip-flop take place.

The graphs in Fig. 10-72 show that oscillation waveforms feature sharp variations, i.e., they are of a relaxation type. The spectrum of such oscillations consists of a great number of high level harmonics which are the major cause why this oscillator was called the multivibrator. The above analysis of circuit operation makes it possible to conclude that the duration of separate processes in a cycle depends upon the time constants of the discharge and charging circuits of capacitors C_1 and C_2 which allows one to change easily their duration within wide ranges. Therefore, the multivibrator circuit with its numerous versions are extensively applied in pulse radio devices.

Extensive use had been also made in pulse radio devices of a *blocking oscillator* which makes it possible to obtain

extremely short and high amplitude pulses whose duration and repetition frequency can be changed comparatively simply and within wide ranges.

One of the possible circuits of the blocking oscillator is illustrated in Fig. 10-73. The primary winding of a transformer having an iron or ferrite core is inserted in the anode circuit; the secondary winding of this transformer is connected to the grid through an automatic bias network consisting of capacitor C and resistor R .

Transformer windings are connected in such a way that a positive potential at the end of the secondary winding coupled with the grid increases as anode current rises. Thus, a strong positive feedback exists between the anode and grid circuits.

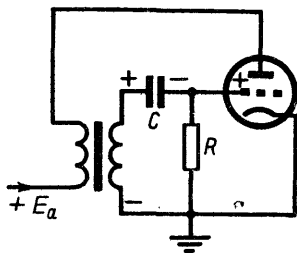


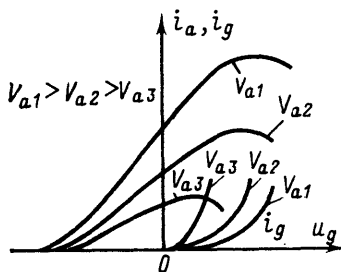
Fig. 10-73. Blocking oscillator

Let us consider how oscillations are produced in such a circuit. Current arises in the anode circuit when the blocking oscillator is connected to a supply source. An increase in this current gives rise to a positive potential at the upper end of the secondary winding. This positive potential is transferred through capacitor C to the grid (the positive charges on the left-hand plate of the capacitor will attract negative charges from the grid to its right-hand plate). An excess of positive charges at the grid will cause a further increase of anode current and this, in turn, will cause a much greater increase of the grid positive potential, etc. This is the way how an avalanche-type rise of anode current which usually lasts less than a hundredth fraction of a microsecond occurs in the circuit.

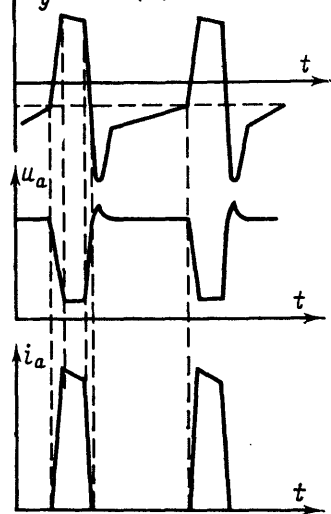
This process stops due to the fact that increase in the grid positive potential u_g leads to the redistribution of electron flow in the valve, i.e., an ever increasing portion of this flow is intercepted by the grid and a lesser portion reaches the anode. As a consequence, valve transconductance is also reduced (Fig. 10-74a). In some cases the grid current may even exceed the anode current. The grid current will charge capacitor C as the anode current increases and also

during a small time interval after the latter stops to increase.

Voltage across the secondary winding disappears as soon as the increase of anode current i_a ceases. As a result, grid



(a)



(b)

Fig. 10-74. Processes in blocking oscillator

(a) anode- and grid-current characteristic curves of triode; (b) currents and voltages in blocking oscillator

voltage starts to drop. Anode current at first is little affected as the valve operating point lies within the region of a low transconductance. The process of capacitor discharge, however, progresses very quickly since the grid current is very high. Therefore, for a short period of time (of the order of microseconds) grid voltage decreases down to such a value that anode current starts to fall off markedly. The decrease in anode current gives rise to negative voltage across the secondary winding which aids a further decrease of anode current. The avalanche-type drop of anode current results in valve cutoff. Simultaneously, the positive voltage across the control grid becomes quickly negative due to a high negative voltage across the secondary winding.

The abrupt decrease in anode current down to zero gives rise to a high short pulse of negative voltage across the secondary winding. On completion of this pulse, the capacitor starts to discharge through the secondary winding and resistor R having a high resistance.

At this stage an input impedance of the cutoff valve is very high (since no grid current flows) so that it does not

shunt the resistor R . Therefore, the capacitor discharges for a long period and the negative voltage across the grid falls for a more or less extended time down to the magnitude at which anode current appears once again and a new momentary pulse is produced.

The graphs of the processes occurring in the circuit (Fig. 10-74*b*) show that oscillations produced by the circuit have the form of short positive and negative pulses of grid (u_g) and anode (u_a) voltages, respectively. Pulse duration is determined basically by the static characteristics of the valve, the magnitude of grid current and the capacitance of the capacitor. The capacitance of the capacitor is easily adjusted. The interval between pulses (i.e., their repetition frequency) may be also adjusted within wide ranges by changing the resistance of resistor R . Such an adjustment makes it possible to change separately the pulse duration and repetition frequency.

Chapter XI. CONVERSION OF ELECTRIC SIGNALS

11-1. CONCEPT OF SIGNAL CONVERSION

The conversion of electric signals occurs in any radio system. In order to use high-frequency oscillations for the transmission of primary signals, it is necessary to find some means of exerting influence on these oscillations by transmitting them in the form of separate pulse trains of definite duration or by changing their amplitude, phase or frequency in accordance with the signal to be transmitted. When receiving these RF signals, it is necessary to restore the primary (original) signal on the basis of variations in one of the parameters of RF signals or those in pulse duration. The processes occurring therewith and associated with the modification of waveforms, frequency or phase will be referred to as *signal conversion*.

The kinds of signal conversion and converter circuits are numerous and varied. We will consider here only the most important and widely used ones. All kinds of conversion may be classified as *linear* and *non-linear*. The conversions of the first type realized in linear circuits obeying Ohm's law may result only in changes of amplitude and phase relations in the complex-signal spectrum. Conversions occurring in non-linear circuits are more complex and result generally in the change of the signal spectrum itself, i.e., they produce such output components which were not present in the input signal. Among linear circuits are those whose parameters are independent of the currents and voltages acting in them. In non-linear circuits, on the contrary, any changes in currents and voltages give rise to changes in their resistances and reactances. As a consequence, voltages and currents

are not proportional to one another and are related by more complex relationships.

The same device may function frequently both as a linear and non-linear element depending upon the method of its use. For instance, if a valve is operated within the portion *AB* of its characteristic curve idealized in the form of a piecewise linear curve (Fig. 11-1), it will function as a linear element. If the same valve is used within the portion *CB*, it will function as a non-linear element and it may be employed for signal conversion.

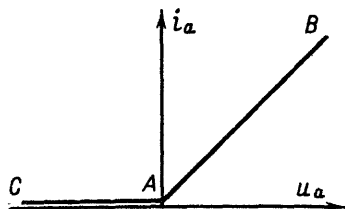


Fig. 11-1. Regions of valve application as linear and non-linear element

Let us familiarize ourselves first with one of the most important kinds of conversion of RF signals, namely, *modulation*. To transmit audio signals by radio, one of the parameters of high-frequency oscillations such as amplitude, frequency or phase is changed in accordance with the signal to be transmitted. This process is called *amplitude*, *frequency* or *phase* modulation, respectively. Amplitude modulation is the simplest and most widely used kind of modulation.

11-2. AMPLITUDE MODULATION

Assume that it is necessary to transmit a simple audio signal of one frequency and constant amplitude. To do this, acoustical (sound) oscillations should be first converted into the electric signal. This problem is generally solved with the aid of a microphone. The AC voltage obtained from the microphone is fed either directly or after amplification to a modulator. This voltage (Fig. 11-2a) obeys the following law

$$u_{au} = V_{au \max} \sin \Omega t \quad (11-1)$$

where Ω is the audio signal frequency.

In case of amplitude modulation (AM) the amplitude of high-frequency oscillations should change in accordance with this law (Fig. 11-2b), i.e., fall off at the negative half-cycle of signal voltage and rise during the positive

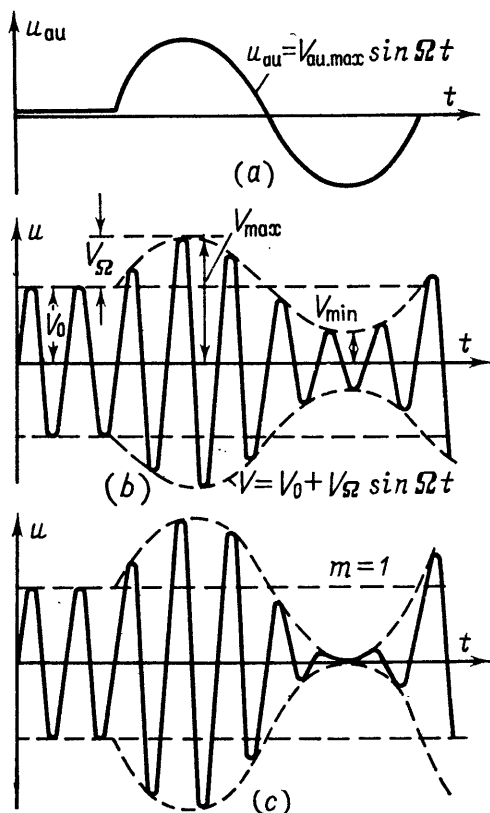


Fig. 11-2. Amplitude-modulated signal

(a) modulating voltage; (b) modulated signal with $m < 1$; (c) modulated signal with $m = 1$

half-cycle. This can be expressed as

$$V = V_0 + V_\Omega \sin \Omega t \quad (11-2)$$

where V is the instantaneous value of RF signal; V_0 , the high-frequency oscillation amplitude (in the absence of modulation), i.e., when the audio signal acting on the microphone is zero, and V_Ω , the maximum deviation of RF signal amplitude from the value V_0 when the audio signal acts on the microphone.

This expression is often written in the form

$$V = V_0 \left(1 + \frac{V_\Omega}{V_0} \sin \Omega t \right) \quad (11-3)$$

A relative value of a maximum change in the amplitude of RF signals is called a *modulation factor*, or *modulation depth* defined as

$$\frac{V_\Omega}{V_0} = m \quad (11-4)$$

This ratio is multiplied by 100 to obtain the modulation percentage

$$\frac{V_\Omega}{V_0} 100 = m\%$$

When $\sin \Omega t = 1$, RF signal amplitude V has its maximum value,

$$V_{\max} = V_0 (1 + m) \quad (11-5)$$

and when $\sin \Omega t = -1$, its minimum value,

$$V_{\min} = V_0 (1 - m) \quad (11-6)$$

By summing up equations (11-5) and (11-6), we obtain

$$V_0 = \frac{1}{2} (V_{\max} + V_{\min}) \quad (11-7)$$

and their difference gives

$$m = \frac{V_{\max} - V_{\min}}{2V_0} \quad (11-8)$$

From equations (11-8) and (11-7) it follows that

$$m = \frac{V_{\max} - V_{\min}}{V_{\max} + V_{\min}} \quad (11-9)$$

It is evident that the maximum modulation factor will be in that case when the RF signal amplitude varies from $V_{\min} = 0$ to $V_{\max} = 2V_0$; in this case, as it follows from equation (11-9), $m = 1$ (or $m = 100\%$). The graph of signals for this case is shown in Fig. 11-2c.

The equation of an amplitude-modulated RF signal may be written in the form

$$u = V \sin \omega t = V_0 (1 + m \sin \Omega t) \sin \omega t \quad (11-10)$$

where ω is the RF signal frequency.

By opening the brackets, we obtain

$$u = V_0 \sin \omega t + V_0 m \sin \Omega t \sin \omega t \quad (11-11)$$

Since

$$\sin \alpha \sin \beta = \frac{1}{2} [\cos (\alpha - \beta) - \cos (\alpha + \beta)]$$

the equation of an AM signal may be presented in another form

$$u = V_0 \sin \omega t + \frac{V_0 m}{2} \cos (\omega - \Omega) t - \frac{V_0 m}{2} \cos (\omega + \Omega) t \quad (11-12)$$

Consequently, the AM signals represent the result of summation of three high-frequency components. The oscillations generated at frequency ω have the maximum amplitude. This frequency, ω , is generally called *carrier frequency*. The two other components have equal amplitudes not exceeding half the amplitude of the carrier-frequency component and the frequencies lying below and above the carrier frequency and differing from it by a value equal to the modulating signal frequency. These frequencies are generally referred to as lower and upper *side frequencies* (or simply sidebands).

The fact that ordinary AM signals consist of three high-frequency components having constant amplitudes may be illustrated graphically. Sidebands produce beats, i.e., the oscillations having the carrier frequency and amplitude envelope which varies at the difference frequency, 2Ω . They are summed up with the carrier-frequency oscillations having a constant amplitude. As the amplitude of beats rises and falls, they increase first the amplitude of the resultant wave, but when the phase is changed by 180° (this happens after the amplitude of beats has passed through zero), they decrease the amplitude of the resultant wave (Fig. 11-3a).

This process may also be illustrated in vector form (Fig. 11-3b). If the vector of carrier wave at the initial instant of time $t=0$ takes a vertical position, then in compliance with expression (11-12) which may be written as

$$u = V_0 \sin \omega t + \frac{V_0 m}{2} \sin [(\omega - \Omega)t - 90^\circ] + \frac{V_0 m}{2} \sin [(\omega + \Omega)t + 90^\circ] \quad (11-13)$$

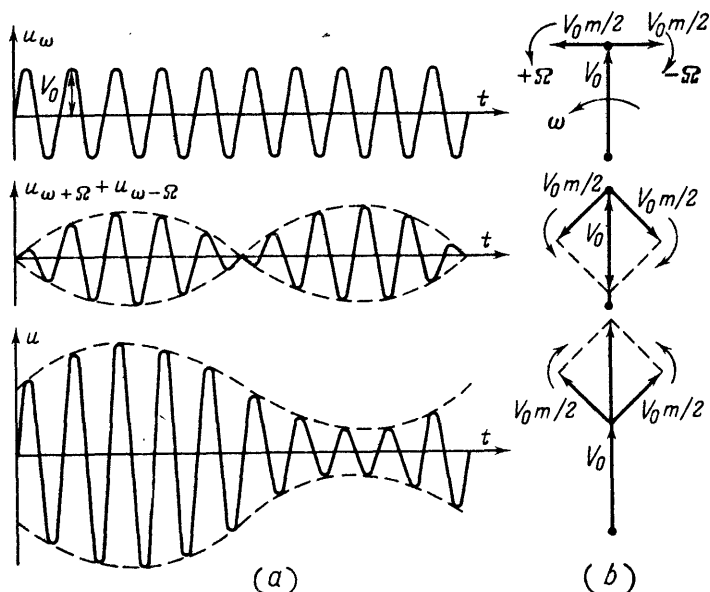


Fig. 11-3. Components of amplitude-modulated signal and their vector representation

the vector of lower sideband will be represented by a horizontal vector directed to the right, whereas the vector of upper sideband will be depicted by the same vector directed to the left.

As the vector of upper sideband rotates counterclockwise at angular frequency $\omega + \Omega$ and the vector of carrier wave performs the same motion at angular frequency ω , the vector of the upper sideband will rotate counterclockwise with respect to it at angular frequency $+\Omega$. At the same time the vector of lower sideband will rotate by the same reason with respect to the vector of carrier wave at constant angular frequency $-\Omega$, i.e., clockwise. It can be easily understood from the vector diagrams shown in Fig. 11-3b that the vector equal to the sum of the vectors of sidebands is always directed along the vector of carrier wave, thus decreasing or increasing alternately its length.

To reproduce well human speech, it is necessary to transmit audio frequencies within a bandwidth from 100 to 3500 Hz, and for music within a bandwidth from 50 to 10,000 Hz. Therefore, transmission of modulated signals generally results in radiation of some spectrum of high frequencies whose limits are determined by the maximum upper and minimum lower side frequencies, $f_0 + F_{\max}$ and $f_0 - F_{\max}$ (Fig. 11-4). Consequently, *in case of amplitude modulation the transmit-*

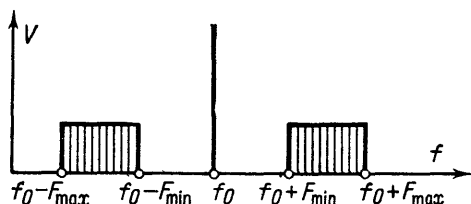


Fig. 11-4. Spectrum of amplitude-modulated signal

ter occupies a certain frequency band whose width is twice the maximum modulation frequency, i.e., $2F_{\max}$.

When a radio station is said to operate at frequency f , by this is meant the carrier frequency f_0 . Two radio stations will always produce mutual interference if their carrier frequencies differ from each other by less than the sum of their maximum modulation frequencies. Therefore, for instance, the carrier frequencies of broadcasting stations differ in some ranges by at least 9 kHz.

To avoid distortions in the transmission and reproduction of radio signals, the transmitter and receiver should possess a fairly wide passband, i.e., sidebands should not be essentially attenuated as compared with the carrier wave. Moreover, it is essential that tuned circuits could be tuned to the carrier frequency as accurately as possible. If these circuits are mistuned, one of the sidebands will be more attenuated than the other sideband and the signal will be heavily distorted. Such distortions occur, for instance, when the receiver is poorly tuned.

In case of amplitude modulation the transmission distance is determined by the level of the sidebands carrying the intelligence. The amplitude (level) of the sidebands is always

less than that of the carrier wave. From expression (11-12) it follows that the amplitude of sidebands is proportional to the modulation factor and, consequently, *the greater the modulation depth of a transmitter, the greater will be the transmission distance.*

AM signals can be obtained by applying high-frequency oscillations and an audio frequency signal simultaneously to the input of a valve tuned amplifier (Fig. 11-5) if the

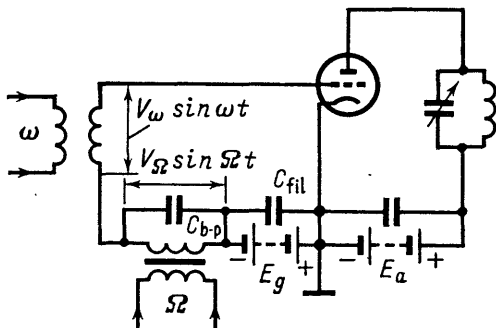


Fig. 11-5. Bias grid modulation

operating point of the valve is selected in the non-linear portion of its anode-grid characteristic curve. In this circuit configuration, referred to as a *grid modulation circuit*, high- and audio-frequency voltages are fed to the valve grid circuit from the secondary windings of the transformers whose primary windings are supplied with oscillations from a high-frequency continuous-wave oscillator and an audio-frequency amplifier, respectively. In addition, the grid circuit is provided with a bias voltage DC source E_g (C-power supply).

As the secondary winding of an audio-frequency transformer having an iron core offers a high impedance for high-frequency currents, it is generally by-passed by capacitor C_{b-p} . The capacitance of this capacitor should be chosen so that it can present a low reactance to high-frequency currents and a high reactance to audio-frequency currents. Since these frequencies differ from one another thousands of times, this requirement is usually met. The bias voltage

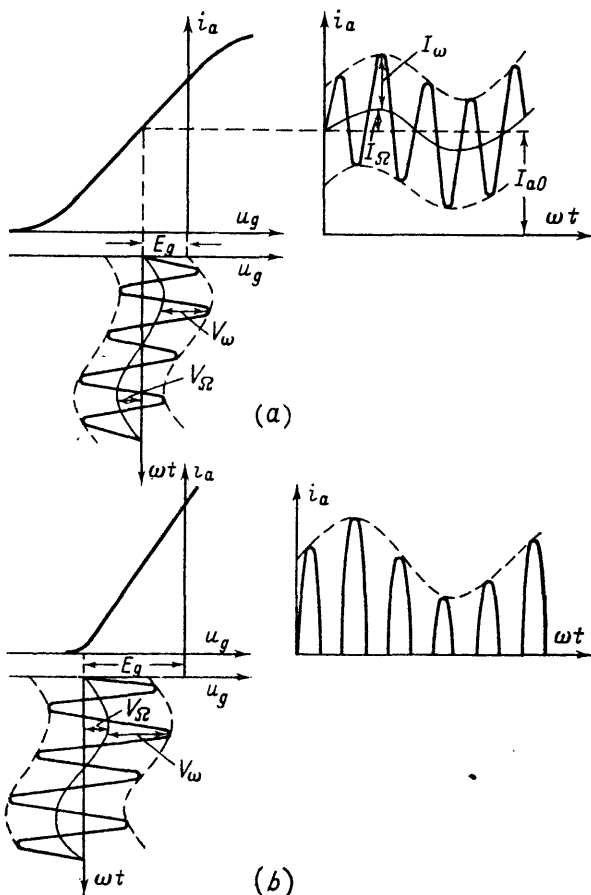


Fig. 11-6. Selecting the valve operating condition in case of bias grid modulation

(a) when operating the valve in linear portion of characteristic curve, no modulation occurs; (b) modulation takes place when the valve is operated in non-linear portion of characteristic curve

source is also blocked by capacitor C_{fil} having a large capacitance so that no high- and audio-frequency currents can pass through it and produce a voltage drop across it.

Let us represent graphically the processes occurring in the circuit considered. To do this, we draw a time scale

under the static characteristic curves of the valve and plot the total voltage acting in the grid circuit as a function of time, $u_g = -E_g + V_\omega \sin \omega t + V_\Omega \sin \Omega t$. Let us consider first the case when only the linear portion of the valve characteristic curve is used. From the graphs shown in Fig. 11-6a it can be easily seen that anode current is simply the sum of direct current I_{a0} , audio-frequency current I_Ω , and constant-amplitude high-frequency current I_ω which produces a constant voltage drop across the tuned circuit connected in the anode circuit (Fig. 11-5). In this case, no modulation occurs in the circuit.

If the non-linear portion of the static characteristic curve (Fig. 11-6b) is used, anode current represents a periodic train of RF pulses whose envelope varies in compliance with the audio-frequency signal being transmitted. It is evident that the higher the amplitude of the current pulses, the higher will be the amplitude of the oscillations excited by them in the tuned circuit. Therefore, the RF signal whose amplitude varies in compliance with the signal being transmitted, i.e., the AM signal is produced in the circuit. This may be also explained as follows. The amplitudes of all the harmonics and, consequently, voltage across the circuit tuned to one or another harmonic change proportionally to the pulse amplitude.

Amplitude modulation may be accomplished by changing any of the supply voltages associated with a tuned amplifier provided that the non-linear portion of the valve characteristic curve is used. In an *anode modulation circuit* (Fig. 11-7a) anode voltage is changed by applying the voltage from a modulation-transformer secondary winding connected in series with the anode-voltage DC source E_a usually called a B power supply (the parallel feed is used to supply the anode circuit of the valve). The amplitude of anode current pulses and, consequently, that of their harmonics vary as the anode voltage is changed. This occurs by two reasons: (1) a decrease in anode voltage results in redistribution of current in the valve in which anode current is reduced at positive grid voltages because of an increase of grid current; (2) the increase in grid current leads, in its turn, to the rise of the bias voltage produced by the automatic bias network, thus decreasing the maximum grid voltage (Fig. 11-7b).

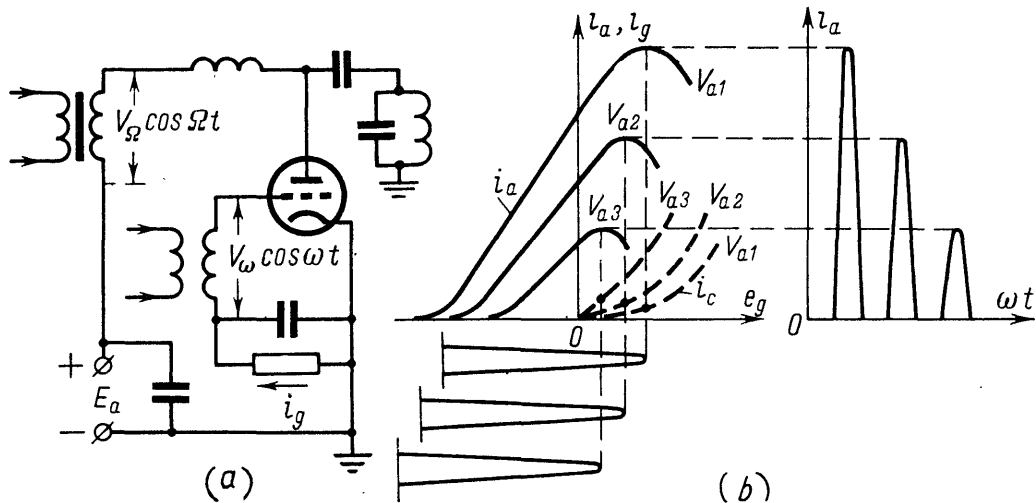


Fig. 11-7. Anode modulation
 (a) schematic diagram; (b) currents and voltages during modulation

The anode modulation makes it possible to obtain a deeper undistorted modulation and higher efficiency than the grid modulation but requires a many times greater level of the modulating signal. Therefore, in practice, both kinds of modulation are used.

A tetrode cannot be used in the anode modulation circuit due to the dynatron effect occurring in such a valve when its anode voltage drops significantly. Therefore, when employing the tetrode it is common practice to use a combined anode-screen modulation in which anode voltage and screen-grid voltage are varied in phase, a decrease in anode voltage is accompanied by a proportional decrease in screen-grid voltage so that no dynatron effect takes place.

A schematic diagram of anode-screen modulation is shown in Fig. 11-8. The screen grid is fed here from an anode voltage supply source through resistor R_{g2} . While flowing through this resistor, screen-grid current produces a voltage drop across it. Consequently, only some part of the modulating voltage acting in the anode circuit is applied to the screen grid. A simultaneous drop in anode voltage and screen-grid voltage results in a redistribution of currents in the valve, namely, anode current and screen-grid current decrease, whereas control-grid current increases. The increase in control-grid current leads to an increase of voltage across the automatic bias network, which, in turn, aids in reducing anode current. The anode-screen modulation circuit employing a tetrode makes it possible to obtain undistorted modulation with great modulation factors.

In the case of pentodes modulation is usually accomplished by applying the modulating voltage to their suppressor screen (Fig. 11-9). Generally, pentodes are operated with a small positive voltage supplied to this grid. If we reduce this voltage or apply a negative voltage, the suppressor-grid field will decrease the anode accelerating field and electron flow will be redistributed in favour of the screen and control grids, so that anode current will decrease. In this circuit (just like in the anode-modulation circuit) an automatic bias network is used to increase bias voltage across the control grid of the pentode. Since the voltage across the suppressor-grid is negative during almost the entire cycle of modulating voltage, the current in its circuit is very small. Therefore,

this modulation scheme requires a very low level of modulating signal power. To obtain 100-per cent modulation, negative voltage across the suppressor-grid should be high enough to neutralize completely the anode accelerating field

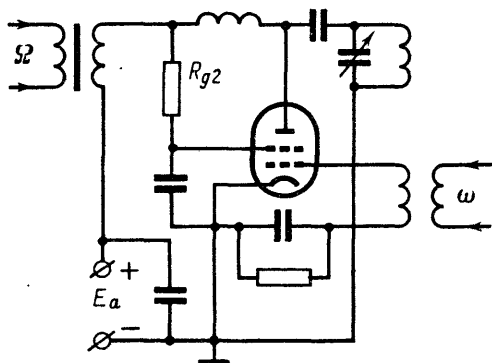


Fig. 11-8. Anode-screen modulation

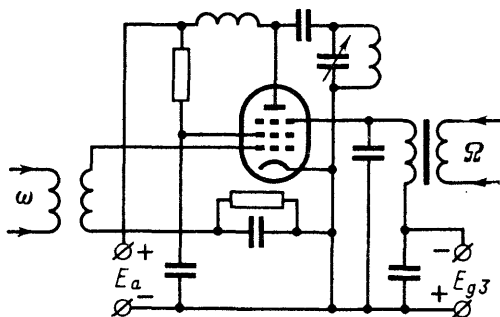


Fig. 11-9. Screen-grid modulation

since only in this case anode current will be fully cut off. Owing to the fact that the suppressor grid is located closer to the cathode than the anode, its effect on the anode current is usually 5 to 8 times stronger than that of the anode. The suppressor-grid modulation also provides undistorted modulation with great modulation factors. Among its drawbacks are relatively high modulating voltages and increased driving power levels caused by appreciable currents in the control-grid circuit.

11-3. FREQUENCY AND PHASE MODULATION

The main disadvantages inherent in amplitude modulation are as follows: (1) the reception of AM signals is severely distorted by man-made and atmospheric interference; (2) the modulator valve is used inefficiently since during the most part of its cycle the modulating voltage is less than the rated voltage of the valve. These disadvantages are eliminated to a large extent when frequency and phase modulation is used. In these cases the amplitude of RF signal remains constant, thus allowing the efficient use of the modulator valve. The constant amplitude of the transmitted signal makes it possible to eliminate in the receiver the noise superimposed on the useful signal without distorting its modulation.

In phase modulation (PM) the phase of the resultant RF signal varies according to the audio signal being transmitted. This may be expressed in the following form

$$\varphi = \omega t + \Delta\varphi \sin \Omega t + \varphi_0 \quad (11-14)$$

where $\Delta\varphi$ is the peak difference between the instantaneous phase angle of the modulated wave and the phase angle of the carrier wave referred to as the *phase deviation*. Here, the phase deviation $\Delta\varphi$ is directly proportional to the intensity of the audio frequency signal (i.e., to the modulating voltage amplitude), $\Delta\varphi = kV_\Omega$, where k is the proportionality factor. The equation of PM signal may be written as

$$a = A_0 \sin (\omega_0 t + \Delta\varphi \sin \Omega t + \varphi_0) \quad (11-15)$$

In the case of frequency modulation (FM), frequency is varied in accordance with the audio frequency signal being transmitted, or

$$\omega = \omega_0 + \Delta\omega \sin \Omega t \quad (11-16)$$

where $\Delta\omega$ is the peak difference between the instantaneous frequency of the modulated wave and its carrier frequency, called *frequency deviation*. As this takes place, frequency deviation is proportional to the intensity of the audio frequency signal (i.e., to the modulating voltage amplitude), $\Delta\omega = kV_\Omega$. In frequency modulation with a sinusoidal modulating wave, the ratio of the frequency deviation to the fre-

quency of the modulating wave is referred to as the *modulation index*. An FM signal as a function of time is shown in Fig. 11-10.

Let us consider the relation between PM and FM signals. The angular frequency of oscillations is equal to the rate of change of phase angle φ . Thus, if the angular frequency varies, we should talk about its instantaneous magnitude. In phase modulation the phase of oscillations varies in ac-

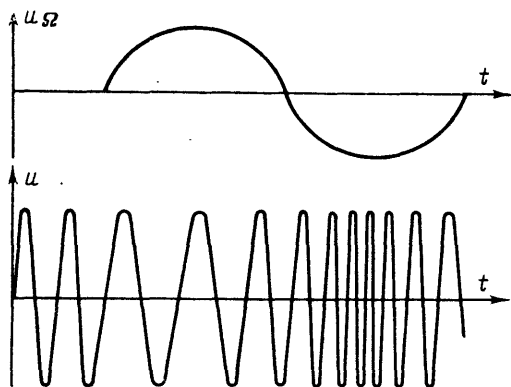


Fig. 11-10. Frequency-modulated signals

cordance with equation (11-14), i.e., it consists of three components, namely, one component varying at a constant rate ω_0 , the other varying by a sinewave law, and the third representing the constant component φ_0 . Therefore, the angular frequency, i.e., the rate of phase change, also consists of three corresponding components. For a sine wave, $u = V_0 \sin \omega t$, the instantaneous rate of change is determined by the expression $\omega V_0 \cos \omega t$ (see Chapter I). Therefore, the rate of change of the second phase component is equal to $\Omega \Delta \varphi \cos \Omega t$. Since the rate of change of the first component (ω_0) is constant and that of the third component is zero, the total rate of phase change, i.e., the angular frequency will be

$$\omega = \omega_0 + \Omega \Delta \varphi \cos \Omega t \quad (11-17)$$

Comparing this expression with equation (11-16) for FM, it is easy to see that they are totally identical in form and,

consequently, FM arises simultaneously with PM and vice versa.

In phase modulation, frequency deviation is defined as

$$\Delta\omega = \Omega\Delta\varphi \quad (11-18)$$

Since $\Delta\varphi$ is proportional to the amplitude of the modulating signal, the frequency deviation proves to be proportional to the intensity of the audio signal but as well as to its frequency, i.e., in phase modulation higher audio frequencies produce a greater frequency deviation than lower audio frequencies. It is quite evident that the reverse situation is observed in frequency modulation, namely, the phase modulation produced by it is characterized by a phase deviation which is inversely proportional to the frequency of the modulating signal. Thus, we may write that

$$\Delta\varphi = \frac{\Delta\omega}{\Omega} = \beta \quad (11-19)$$

i.e., the lower audio frequencies produce deeper modulation than higher audio frequencies. All the above indicates that FM signals may be converted into PM signals and vice versa.

Theoretical analysis of FM signals shows that they may be presented as an infinite series of harmonics (sine waves) whose frequencies differ from one another by the frequency of the modulating signal. Thus, an FM signal may be represented in the form

$$\begin{aligned} u = & V_0(\beta) \sin \omega_0 t + V_1(\beta) [\sin(\omega_0 + \Omega)t - \\ & - \sin(\omega_0 - \Omega)t] + V_2(\beta) [\sin(\omega_0 + 2\Omega)t + \\ & + \sin(\omega_0 + 2\Omega)t] + \dots \\ & \dots + V_n(\beta) [\sin(\omega_0 - n\Omega)t + (-1)^n \sin(\omega_0 + n\Omega)t] + \dots \end{aligned} \quad (11-20)$$

In other words, FM and PM are characterized by an infinite spectrum of sidebands.

It would seem that the transmission and reception of such signals are not feasible. Analysis shows, however, that the amplitudes of sidebands decrease rather quickly with their increasing number. This decrease occurs the quicker the less the modulation index. All the sideband components whose numbers exceed the modulation index, expressed in radians,

are found to have amplitudes less than 10-15 per cent of the carrier wave and contain not more than 1 to 2 per cent of the transmitted signal energy. In practice, therefore, they may be disregarded.

When the modulation index $\Delta\phi \ll 1$, all sideband components whose number exceeds unity prove to be very weak and may be neglected. In this case, as seen from expression (11-20), the spectrum consists practically of carrier frequency and two side frequencies, i.e., it contains the same components as in the case of amplitude modulation. Such a frequency or phase modulation is called *narrow-band modulation*.

Interference (noise) exerts less effect on the reception of the FM signals whose modulation index $\Delta\phi \gg 1$. The number of required sideband components is limited by

$$n = \Delta\phi = \frac{\Delta\omega}{\Omega} \quad (11-21)$$

In this case, spectrum width will be

$$2\Delta\omega_{\max} = 2n\Omega = 2\Delta\omega \quad (11-22)$$

i.e., for *wideband* frequency modulation with $\Delta\phi > 1$, the signal spectrum width used is twice the value of frequency deviation.

The spectrum proves to be wide since in this type of modulation the frequency deviation usually exceeds many times the modulating signal frequency. For this reason FM is not used in the range of long, medium and short waves. Practically, FM is utilized in transmitters operating at metre and shorter waves.

Frequency modulation is frequently accomplished with the aid of the electron valves which being connected in a special circuit configuration serve as variable reactances; they are connected in parallel to the tuned circuit of a self-excited oscillator (such valves are called reactance valves). If we change their equivalent reactance in accordance with the audio-frequency signal being transmitted, then the tuned-circuit resonance frequency and, consequently, the frequency of the signal being generated will also change in accordance with the modulating voltage.

One of the FM schemes employing a reactance valve is shown in Fig. 11-11a. A valve with a high amplification factor (usually a pentode) is connected in parallel to an oscillator tuned circuit so that the AC voltage of the tuned circuit is also applied to the anode of the valve. The latter will function as a reactance if the current through it is shifted by 90° with respect to the AC anode voltage. To produce such

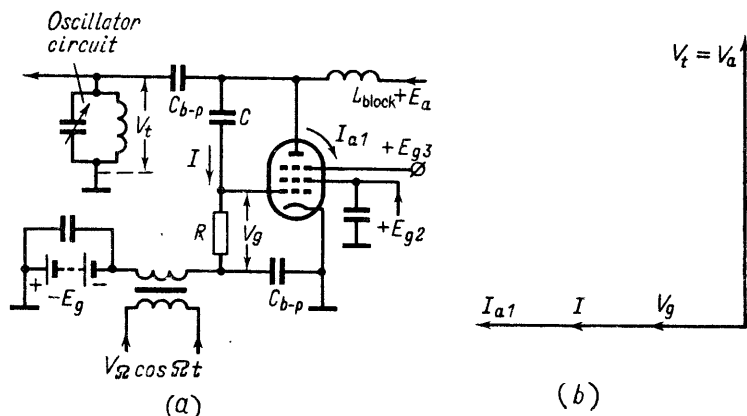


Fig. 11-11. Frequency modulation with the aid of reactance valve
(a) schematic diagram; (b) vector diagram

a shift, voltage is fed from the tuned circuit to the control grid through a phase shifter consisting of capacitor C and resistor R . The parameters of the phase shifter are selected so that it can shunt to the minimum extent the tuned circuit, and the capacitor reactance is many times as great as the resistance of the resistor, i.e.,

$$\frac{1}{\omega C} \gg R \quad (11-23)$$

The by-pass capacitors C_{b-p} connected in the circuit should offer very low reactance to RF currents so that they may be considered as, practically, a short circuit for these currents.

If the vector, $V_t = V_a$, representing the voltage across the tuned circuit takes a vertical position (Fig. 11-11b) at some instant of time, then the vector, I , representing the current in the RC circuit having a capacitive reactance leads vector

V_a by 90° and, consequently, is in a horizontal position. This current is

$$I = \frac{V_a}{\sqrt{R^2 + \left(\frac{1}{\omega C}\right)^2}} \approx V_a \omega C \quad (11-24)$$

It produces across the resistor R the voltage

$$V_g = IR = V_a \omega CR \quad (11-25)$$

which is in phase with current I and presents an AC voltage applied to the control grid.

Owing to the fact that the valve possesses a very high amplification factor, its anode current is determined essentially by the grid voltage only. The amplitude of the first harmonic of anode current is given by

$$I_{a1} = S_{av} V_g = V_a \omega CR S_{av} \quad (11-26)$$

where S_{av} is the average transconductance of the valve. The vector of the anode current first harmonic has the same direction as that of the grid voltage.

The vector diagram shows that the first harmonic of anode current leads the anode voltage by 90° . In other words, the valve behaves as a conventional capacitor. Its equivalent reactance is determined by

$$X_{eq} = \frac{V_a}{I_{a1}} = \frac{1}{\omega C_{eq}} = \frac{1}{\omega CR S_{av}} \quad (11-27)$$

then the capacitance of the equivalent capacitor will be

$$C_{eq} = S_{av} CR \quad (11-28)$$

Modulation is accomplished by varying the capacitance of this capacitor connected in parallel to the tuned circuit. To do this, average transconductance of the valve is varied by changing the bias voltage (transconductance remains constant and no modulation occurs when the valve is operated in the linear portion of its static characteristic curves). To avoid distortions during modulation, it is necessary to select a non-linear portion of the static characteristic curve (Fig. 11-12) so that the transconductance varies proportionally to the voltage across the control grid. To change the bias in accordance with the audio frequency signal, the secondary winding of the transformer is connected into the

control-grid circuit in series with a bias DC source. Signal voltage is applied to the primary winding of this transformer.

Some other reactance-valve circuit configurations may also be used but their operating principle does not differ in general from that described above.

In principle, the PM signal may be obtained by summing up two RF signals shifted in phase by 90° , one of which is amplitude-modulated (Fig. 11-13a). In this case, the phase of the resultant vector varies according to the amplitude of the AM signal.

Large values of phase deviation, however, cannot be obtained by this means because of poor proportionality between the resultant signal phase and the modulated signal amplitude. The desired phase variation may also be obtained by summing up a constant-amplitude signal with the signal shifted in phase by 90° and having a varying amplitude if the latter changes its phase by 180° when passing through zero. The result of vector summation is shown for this case in Fig. 11-13b.

We encountered the signals of this type when analyzing the composition of AM signals—they are produced when

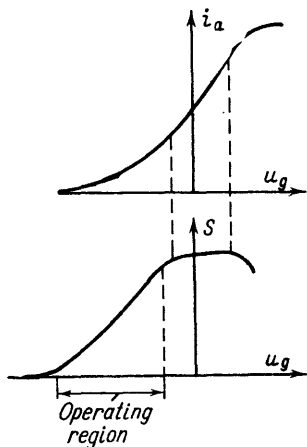


Fig. 11-12. Operating region of reactance-valve characteristic curves

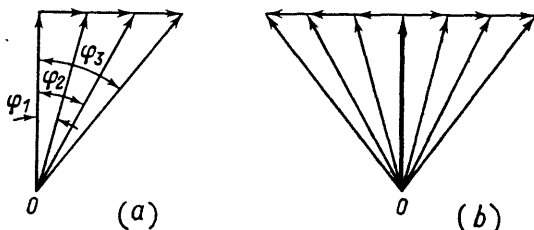


Fig. 11-13. Obtaining the phase-modulated signals

(a) by summation of high-frequency oscillation and AM signal; (b) by summation of high-frequency oscillations and beats produced by sideband AM components

two sidebands beat. How can one separate beat-frequency oscillations? Obviously, for this it is sufficient to suppress the carrier wave in the composition of the AM signal. This may be accomplished, for instance, by combining two grid-modulation oscillators modulated 180° out of phase and operating on a common load. Such a circuit (Fig. 11-14) is referred to as a *balanced modulator* since modulating voltages are applied in it from transformer $Tr2$ to the control

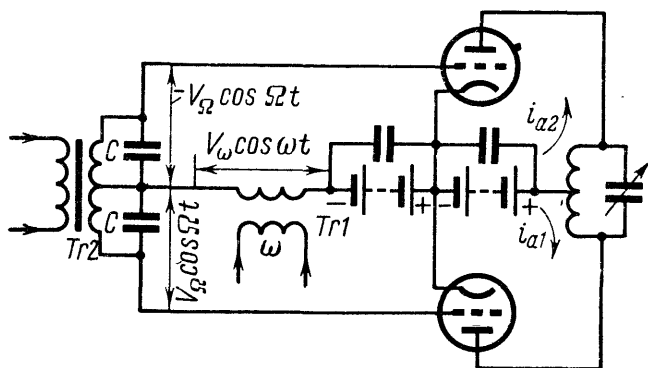


Fig. 11-14. Balanced modulator

grids 180° out of phase and the valves are alternately made conductive or non-conductive. RF voltage is applied to the grids in phase. It is fed from transformer $Tr1$ through blocking capacitors of small capacitance. Therefore, if the anode current in one valve for the case of modulated signal is given by

$$i_{a1} = I_a (1 + m \sin \Omega t) \sin \omega t$$

then for the second valve it will be

$$i_{a2} = I_a (1 - m \sin \Omega t) \sin \omega t$$

Both these currents flow through the common load in opposite directions. Consequently, the voltages produced by them in the circuit are subtracted. Thus, we may write that

$$u_t = i_{a1} R_{eq} - i_{a2} R_{eq} = 2I_a R_{eq} m \sin \Omega t \sin \omega t \quad (11-29)$$

The resultant voltage represents the beat-frequency oscillations produced by the sidebands (see formula (11-11)).

In a transmitter in which phase modulation is accomplished with the aid of the balanced modulator (Fig. 11-15) the output signal of the latter is fed through a phase-shifter which displaces its phase by 90° to the mixer where it is summed up with the constant-amplitude signal supplied from a self-excited oscillator. This circuit may also be used for frequency modulation. For this purpose, it is sufficient for the gain of the audio signal amplifier to be inversely proportional to frequency.

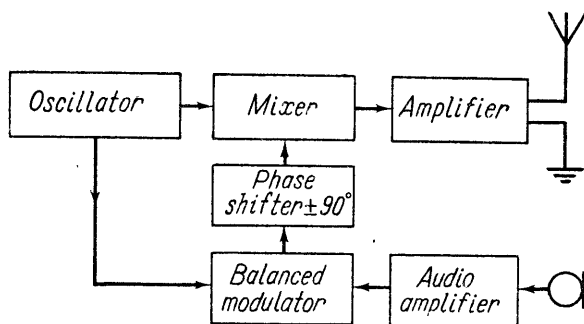


Fig. 11-15. Block diagram of phase-modulated transmitter

For phase modulation the frequency deviation is defined as

$$\Delta\omega = k_1 \Omega V_\Omega$$

If the voltage taken from the amplifier is

$$V_\Omega = \frac{k_2 P_s}{\Omega}$$

where P_s is the sound pressure exerted on the microphone, then

$$\Delta\omega = k_1 k_2 P_s$$

where k_1 and k_2 are proportionality factors.

Consequently, the frequency deviation becomes proportional to the sound pressure and does not depend upon the modulation frequency.

The method used for obtaining frequency-modulated signals from phase-modulated ones is called an *indirect method* of frequency modulation. One of the principal advantages

of this method over the method of direct modulation is the absence of direct influence on the self-excited oscillator, which makes it possible to provide a highly stable center frequency. The indirect method of frequency modulation involves, however, appreciable circuit complication.

The values of frequency deviation obtained with the aid of reactance-valve circuits are not large amounting usually to tenth fractions of a per cent. At the same time in a

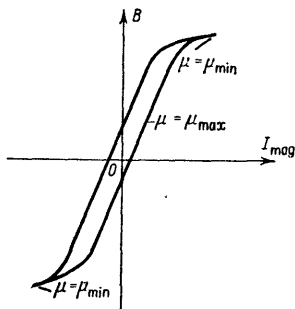


Fig. 11-16. Hysteresis curve for ferrite core

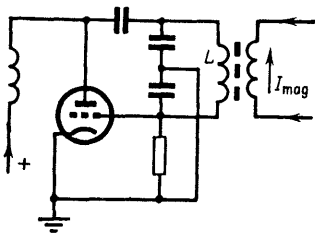


Fig. 11-17. Oscillator with ferrite-core tuning

number of practical applications it is necessary to tune oscillators by varying voltage in much wider limits. At present this problem is solved by several methods. The most simplest is electronic tuning accomplished with the aid of ferrite variometers and crystal diodes.

The inductance of magnetic-cored coils is proportional to the core permeability, μ . In the condition of saturation $\mu = \mu_{\min}$, whereas in the linear portion of a magnetization curve $\mu = \mu_{\max}$ (Fig. 11-16). Magnetodielectrics, especially ferrites, are widely used in the range of radio frequencies, including centimetre waves. They display a very low electrical conductivity (10^{10} - 10^{12} times less than metals) possessing at the same time a very high permeability amounting to several hundreds. Owing to these properties, AC magnetic fields hardly induce eddy currents in them and, therefore, heat losses in ferrite cores are negligible.

If we use a ferrite-cored coil fitted with an additional bias winding (Fig. 11-17) in the tuned circuit of a self-ex-

cited oscillator, which determines its operating frequency, then variation of current I_{mag} in the bias winding will lead to a change in the inductance of coil L and, consequently, to a change in the operating frequency. Such simple circuits allow the frequency of the signal being generated to be varied over a wide range. Unfortunately, in these circuits frequency is very critical to temperature variations and the non-linearity depends on the magnetizing current.

Lesser ranges of frequency variation are obtained in the oscillators whose circuit parameters are varied by changing the $P-N$ junction capacitance of semiconductor devices. It is well known that the capacitance of a cutoff $P-N$ junction may vary, on the average, from units to dozens of picofarads depending on the voltage applied to it. If a crystal diode D cut off by voltage E from a DC supply source (Fig. 11-18) is connected in parallel to the tuned circuit of a self-excited oscillator, the frequency of the signal being generated can be varied by an AC source connected in series with the DC supply source. In such a circuit frequency can be varied by dozens of per cent. The application of this very simple scheme of frequency modulation is somewhat limited because of its non-linearity and strong temperature dependence.

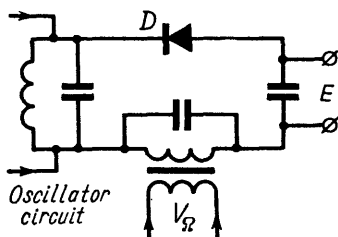


Fig. 11-18. Oscillator with electronic tuning

11-4. PULSE MODULATION

Assume that the transmission of AM signals (Fig. 11-19a) is suddenly interrupted at some instant of time t_1 and then resumed once again at the instant t_2 after a very small time interval Δt . It is quite evident that the operator at the receiving end will not perceive this interruption because a human ear possesses some time constant, i.e., the ability of storing up separate momentary stimuli, i.e., it reacts on their average accumulated value (a similar ability of a human eye allows one to see a continuous motion in cinema rather than separate frames). The number of such interruptions during

transmission may be extremely great without a noticeable deterioration in the quality of signal reproduction at the receiver output.

The intermittent nature of the transmitted signal may be made imperceivable to a great extent if the receiver output is, provided with a filter equalizing short-term pulsations

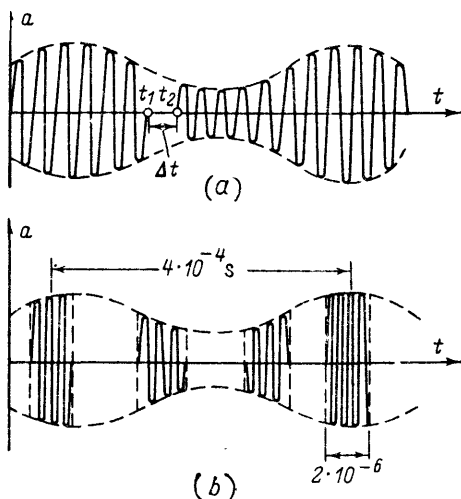


Fig. 11-19. Signal waveforms in pulse-amplitude modulation

of voltage and acting just like the RC filter of an automatic bias network.

Theory and experiments show that to obtain satisfactory reproduction of a received audio signal it is sufficient to transmit the audio-frequency signal at least 3 or 4 times per its cycle during short-term intervals of the order of some microseconds. If we assume that the highest frequency of an audio signal to be transmitted does not exceed 2500 to 3500 Hz, then it can be transmitted in the form of microsecond pulses with a repetition frequency of 8 to 10 kHz.

The amplitude of these pulses is proportional to the amplitude of the useful signal at the instant of radiation. Intervals between the pulses may be hundreds of times greater than the pulse duration (the duration of these pulses in

Fig. 11-19b is heavily enlarged). This makes it possible to use time intervals between the pulses for transmitting pulses of another audio-frequency signal with the aid of the same transmitter. At the receiving end both signals may be received separately due to the fact that they are transmitted at different time instants or, as it is said, are *divided in time*. This kind of modulation is called *pulse-amplitude modulation*, abbreviated PAM.

Thus, the pulse modulation provides a means for simultaneous transmission of several independent signals with the aid of a single transmitter. Such a method of equipment application is known as *multichannel communication*. This is the most important advantage of pulse modulation. Among drawbacks of the pulse-amplitude modulation is low noise immunity of PAM signals just like in the case of ordinary amplitude modulation.

It should be also noted that in the spectrum of short-pulse signals the amplitude of a sideband falls off very slowly with the increasing number of sideband component and the less the pulse duration, the slower this occurs. Therefore, satisfactory reproduction of a pulse signal may be accomplished only with the aid of a receiver having a very wide bandwidth. If, for instance, the pulse duration is equal to $1\mu\text{s}$, the receiver bandwidth should be at least

$$2\Delta f = \frac{1}{1 \times 10^{-6}} = 1 \text{ MHz}$$

This bandwidth exceeds that of the AM receiver by hundreds of times and by several times that of the FM receiver. Therefore, pulse modulation cannot be used in the range of medium, long and short waves.

This kind of modulation has found extensive use in microwave relay systems. It can be accomplished by performing ordinary grid modulation and applying anode voltage in the form of a periodic pulse train (Fig. 11-20). The radio-frequency pulses modulated in amplitude will arise at the circuit output as a result of the simultaneous action of two modulating voltages (and the pulsed voltage).

A much higher noise immunity is possessed by communication links using other kinds of pulse modulation. Historically, the method of *pulse-duration modulation* (abbreviated

PDM) suggested by A. N. Shchukin in 1927 was the first kind of pulse modulation. The PDM signal waveform is shown in Fig. 11-21. The RF pulses are radiated at identical time intervals Δt with duration of each pulse being determined by the amplitude of the modulating signal at the in-

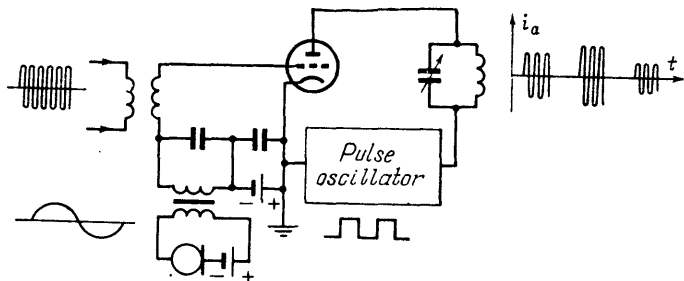


Fig. 11-20. Amplitude-pulse modulation

stant of radiation. In the case of pulse-duration modulation the amplitude of pulses remains constant; this makes it possible to clip in the receiver the interference distorting the amplitudes of the pulses being received. Among disad-

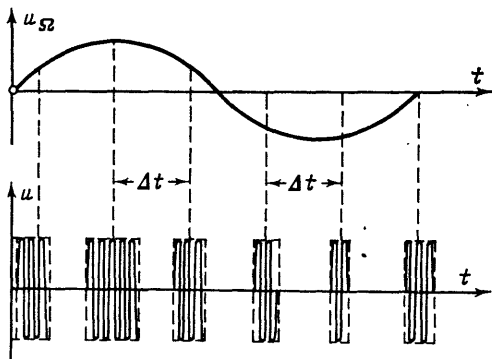


Fig. 11-21. Signal waveforms in pulse-duration modulation

vantages of pulse-duration modulation is the necessity of selecting the bandwidth with respect to the shortest pulse, thus, for all longer pulses it proves to be too wide. However, the wider the receiver bandwidth the greater the amount of interference passes in it.

A more effective use of the receiver passband is obtained with *pulse-position modulation* (also called *pulse-phase modulation*), abbreviated PPM. This kind of modulation is accomplished by sending pulses of equal duration and constant amplitude. The modulation law determines the instant of pulse sending, i.e., the position in time of a pulse is varied (Fig. 11-22). In the absence of modulating voltage pulses

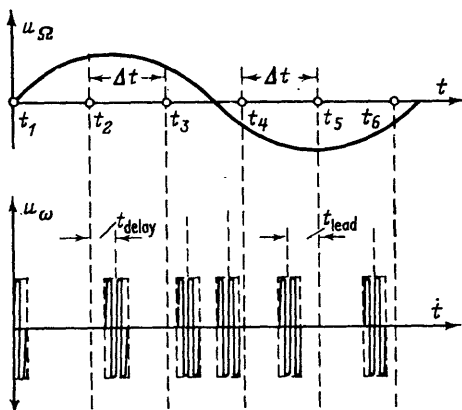


Fig. 11-22. Signal waveforms in pulse-position modulation

are sent at time instants $t_1, t_2, \dots, t_n$ spaced equally at time intervals Δt . If the signal to be transmitted is applied, the instant of radiation of each pulse is delayed with respect to the instants indicated by time intervals t_{delay} proportional to the magnitude of a positive modulating voltage acting at the given moment of time or is advanced (time intervals t_{lead}) proportional to the magnitude of a negative modulating voltage.

The high noise immunity of the communication link and effective use of receiver bandwidth are responsible for the wide practical application of pulse-position modulation. Still other methods of accomplishing the pulse modulation featuring both advantages and disadvantages exist, their description, however, is beyond the scope of this book.

The telegraph mode is similar to pulse modulation by the nature of transmitter application. The use of the Morse

code and the transmission of intelligences coded by this technique were historically the first form of controlling high-frequency oscillations in radio transmitters.

Letters are transmitted by the Morse code in the form of different combinations of long and short signals (dots and "dashes"). Each letter consists of one to five such signals spaced by time intervals during which the telegraph transmitter is blocked. By this means, the transmission of telegraph signals (so-called *telegraph manipulation*, or *keying*) consists

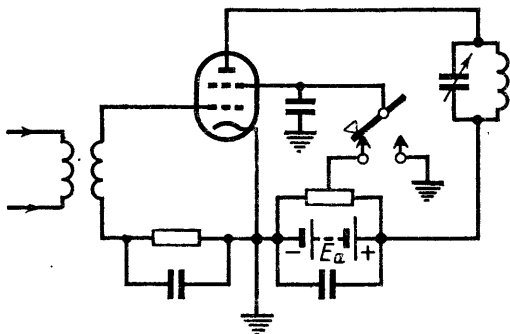


Fig. 11-23. Telegraph manipulation with telegraph key connected into screen-grid circuit

actually in blocking and actuating the transmitter at definite instants of time. Such an operation of the transmitter may be accomplished by different methods, for instance, by connecting a telegraph key (Fig. 11-23) into the supply circuit of the screen grid of an oscillator valve. When the key is pushed down, a normal positive voltage is applied to the screen grid from a voltage divider, which starts the oscillator. When the key is released, the screen grid is connected to the cathode, so that the oscillator is blocked.

In hand-operated telegraph transmitters transmission speeds of 20 to 25 words per minute are attainable whereas modern high-speed telegraph equipment provides transmission speeds of 300 to 500 words per minute. It is easy to compute that at the speeds indicated the duration of telegraph signals amounts to the tenths or thousandths fractions of a second, which is thousands of times greater than the dura-

tion of pulse-modulated signals. Therefore, the receiver bandwidth required for the adequate reproduction of telegraph signals may be taken accordingly thousands of times less than in the pulse-modulation methods described above.

The narrow bandwidth of the order of hundreds of hertz or some kilohertz is one of the most important features of telegraph transmitters and receivers. The effective use of transmitter power and the narrow bandwidth of telegraph signals results in that the transmission distance in telegraph mode is much greater than that obtained in telephone mode with the aid of the same transmitter.

11-5. DETECTION OF RADIO SIGNALS

Detection is the second most important form of conversion of electric signals in radio equipment. Incoming RF signals modulated in amplitude, frequency, phase or having the form of pulses have to be converted in the receiver into the original signals, i.e., into the audio-frequency signals which can be perceived by a human being or registered by instruments. This process of conversion is called *detection*. Just like any signal conversion, detection is feasible only when a non-linear element is used.

AM signals (Fig. 11-24) are the ones most easy to detect. If RF signals modulated in amplitude are applied to a non-linear element, a *detector*, displaying unilateral conductivity (a vacuum or crystal diode may be used as such an element), current in its circuit assumes the form of sinewave pulses whose amplitude varies proportionally to the intensity of the audio-frequency signal.

All the harmonic components of a periodic pulse train and its DC component which is the average value of current for the pulse repetition period are proportional to the pulse amplitude. Consequently, the DC component of the detector current, i_D , varies proportionally to the signal voltage which has modulated high-frequency oscillations in the transmitter. It is filtered out from RF components with the aid of a filter consisting usually of resistor R and capacitor C , having a small capacitance. RF components pass through the capacitor without producing an appreciable voltage across it. This capacitor is also used in order that the entire RF voltage

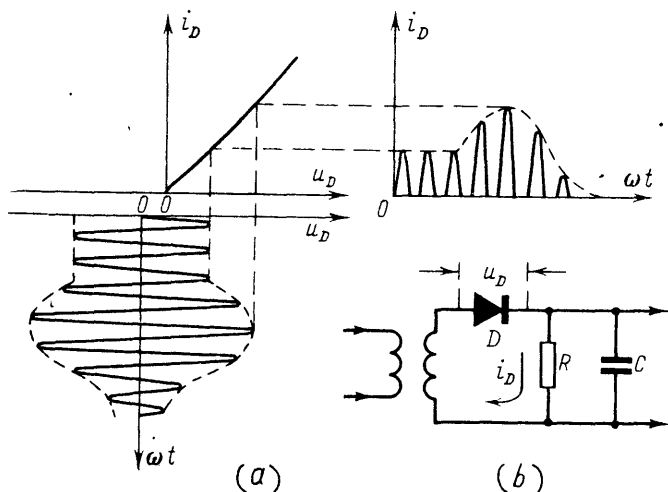


Fig. 11-24. AM-signal detection

(a) processes in detector circuit; (b) detector schematic diagram

being detected is fed to the diode D . (In the absence of the capacitor part of this voltage would drop across resistor R .) The AF current component passing through resistor R produces a voltage across it, which is then fed to the succeeding circuit.

The appearance of audio-frequency voltage across the filter may be explained without resorting to the concept of harmonic current components. Current pulses flowing through resistor R , produce across it a voltage drop which charges the capacitor. For the period between the pulses the capacitor discharges only partially through the resistor. As a result, voltage across the resistor does not disappear completely during the interval between the pulses. Each new pulse imparts an additional charge to the capacitor. Thus, some average voltage varying proportionally to the pulse amplitude is produced across the capacitor.

When detecting FM signals, the variations of instantaneous frequency may be converted first to the high-frequency current variations, i.e., FM-AM conversion is accomplished. This is achieved by feeding the frequency modulated current to the circuit having natural frequency f_{t0} and detuned with respect

to the transmitter centre frequency f_c (Fig. 11-25). In this case transmitter frequency variations to either side result in amplitude variations of high-frequency oscillations in the tuned circuit. By this means, the FM signal is converted to the AM signal. The latter may be detected by the method described above.

In practice, however, use is made of the circuits which make it possible to convert FM signals directly to audio-

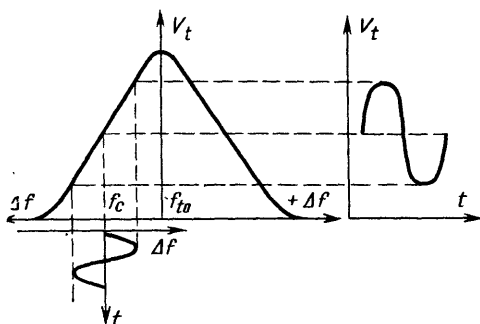


Fig. 11-25. Converting FM signals into AM signals with the aid of detuned circuit

frequency signals. Such circuits are known as *frequency discriminators*.

A schematic diagram of the most extensively used discriminator is illustrated in Fig. 11-26a. The FM signal is taken from circuit 1 tuned to the signal carrier frequency. Circuit 2 is tuned to the same frequency. Both circuits are coupled inductively and capacitively (through coupling capacitor C_m). Inductance of a blocking choke, L_{block} , is chosen so great that it proves to be practically open-circuited for RF currents.

RF voltage is fed to the two diodes connected in the push-pull circuit and loaded with an RC filter whose capacitors C_1 and C_2 present practically a short circuit for RF currents. Taking this into account and inspecting the schematic diagram in Fig. 11-26, we may conclude that the sum of two RF voltages is applied to diode D1, namely, the voltage across circuit 1 and the voltage across the upper (in the diagram) half of the coil in circuit 2, i.e., $V_{D1} = V_{t1} + V_1$, and voltage

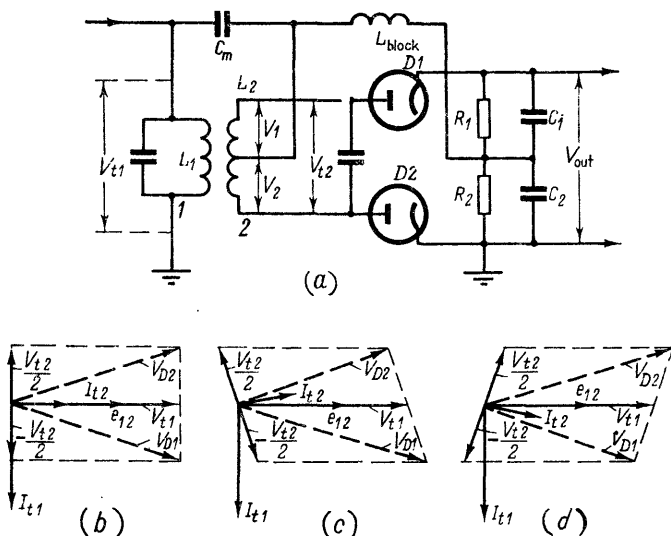


Fig. 11-26. Frequency discriminator

(a) schematic diagram; (b), (c), (d) vector diagrams for different frequencies

$V_{D2} = V_{t1} + V_2$ is applied to the diode D_2 . As these voltages are different in phase, their sums should be considered as vector quantities.

Let us consider first how the circuit will operate when only the carrier frequency to which both circuits are tuned is received. Assume that at some instant the vector of the voltage V_{t1} takes a horizontal position (Fig. 11-26b). Current I_{t1} flowing through coil L_1 lags behind voltage V_{t1} by 90° and EMF e_{12} induced by this current in circuit 2 leads it by 90° in compliance with expression (3-1). In the tuned circuit 2 current I_{t2} will be in phase with this EMF and voltage V_{t2} produced by it across coil L_2 will lead current I_{t2} by 90° . In both diodes the halves of this voltage are summed up with voltage V_{t1} in opposite phases as shown in Fig. 11-26b. Absolute values of the vectors of resultant voltages across the diodes are identical in this case and the sum of voltage drops produced by rectified currents across the load is zero.

The situation will change if the frequency of the oscillations being generated is reduced as a result of modulation. In this

case, the vector diagrams will change as follows: current I_{t2} will lead EMF e_{12} and voltage vector V_{t2} will turn accordingly. As a consequence, the voltage across diode D_1 will become greater than that across diode D_2 (Fig. 11-26c), therefore, the voltage drop produced by the rectified current across resistor R_2 will be greater than that across resistor R_1 and the voltage $V_{out} \neq 0$ will appear at the discriminator output.

By plotting similarly a vector diagram for a case when $f > f_c$ (Fig. 11-26d), it is easy to see that the sign of the output voltage will be reversed. The difference of the voltages across the diodes and, consequently, the output voltage rise proportionally to the deviation of instantaneous frequency from the carrier frequency. This makes it possible to convert FM signals directly to AF signals.

Detection of PM signals may be accomplished in a similar way except that frequency correction should be provided in the audio-frequency amplifier (i.e., the circuit should be designed so that its gain is inversely proportional to the frequency). In the absence of frequency correction, higher frequencies will be reproduced with greater intensity than lower frequencies.

Detection of PAM and PDM signals is accomplished similarly to that of ordinary AM signals.

When detecting the PPM signals, the train of detected pulses following one another at different time intervals is converted to the train of PDM pulses from which the audio-frequency voltage is then obtained.

11-6. FREQUENCY CONVERTERS

In some instances it proves to be necessary to convert the signals of one frequency to those of another frequency which is higher or lower. Frequency is increased an integral number of times with the aid of *frequency multipliers*, which are tuned amplifiers whose load circuits are tuned to one of the higher harmonics of the driving voltage.

A somewhat more complicated process of frequency conversion is used when the frequency of the signals being converted is not an exact integral multiple of the output frequency. Such frequency conversion we have met during the process of modulation. As this takes place, owing to the simultaneous

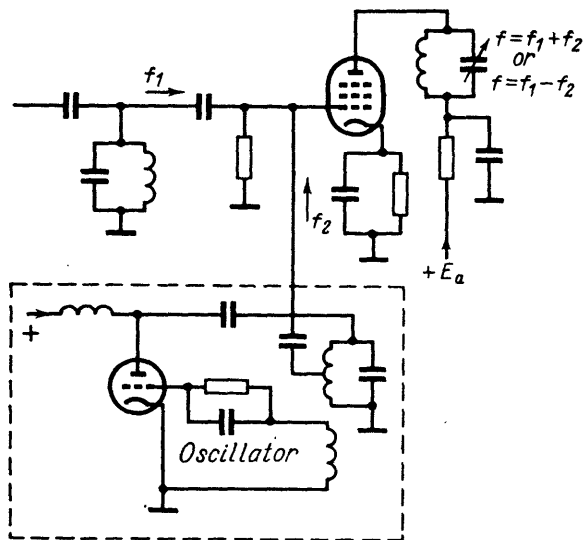


Fig. 11-27. Single-grid frequency converter

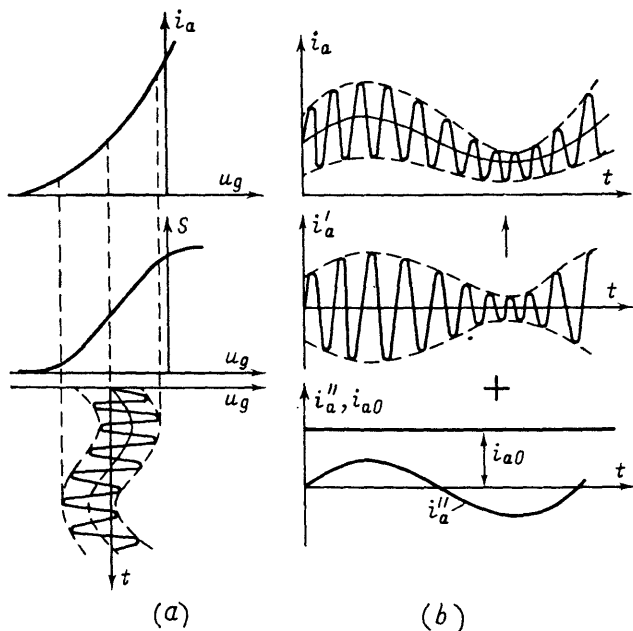


Fig. 11-28. Processes in frequency converter

action of two voltages of different frequencies (carrier frequency and modulation frequency) on the valve operating in the non-linear portion of its characteristic curve, sideband components with frequencies equal to the difference and sum of frequencies of the voltages acting on the valve appear at the valve output in addition to the carrier wave.

A similar method is used for frequency conversion, namely, two voltages of different frequencies are applied simultaneously to a non-linear element which may be a crystal or vacuum diode, triode or a multigrid valve. A schematic diagram of a single-grid frequency converter is shown in Fig. 11-27.

Operating principles of the converter may be explained as follows. The transconductance S of the valve increases and decreases alternately under the action of one of the voltages, for instance, the voltage having a lower frequency. Therefore, the amplitude of the higher frequency signal produced in the anode circuit, which is proportional to the transconductance, will increase and decrease during positive and negative half-cycles, respectively (Fig. 11-28a). The anode-circuit current i_a (Fig. 11-28b) is the sum of three components, namely, DC current i_{a0} , lower frequency signal current i_a'' and higher frequency signal current i_a' having a variable amplitude. The last AM current represents the result of summation of oscillations of the higher frequency and the beats of two new high-frequency signals. Their frequency may be determined proceeding from the properties of waveforms arising during beating. The frequency of amplitude-envelope variations may be determined from formula (10-60),

$$F = \frac{f_1 - f_2}{2}$$

and the frequency of the high-frequency signal from the expression

$$f = \frac{f_1 + f_2}{2}$$

Summing up and subtracting the last two equations, we obtain the frequencies of the oscillations which give rise to the beats, these are

$$f_1 = f + F \quad (11-30)$$

$$f_2 = f - F \quad (11-31)$$

Thus, the currents of the difference and sum frequencies have appeared in the anode circuit. For this reason, frequency converters are also called *mixers* (this term is widely used in technical literature). The signal of the desired frequency can be obtained at the output of the frequency converter by connecting into its anode circuit the circuit tuned to the sum or difference frequency.

It might be well to point out the difference between the process of frequency conversion and that of modulation.

As a rule, modulation is accomplished when the oscillator is operated at the cutoff angle $\theta < 180^\circ$, which is necessary to provide high efficiency. In turn, frequency conversion is usually carried out in low-power devices where high efficiency is not of great importance. One of the most important requirements in these circuits is the minimum content in the anode tuned circuit of combination frequencies of the type

$nf_1 \pm mf_2$ and higher harmonics, which is obtained by operating the oscillator without anode current cutoff, i.e. at the cutoff angle $\theta = 180^\circ$. In addition, during modulation the modulating frequency should be many times less than that of the signal being modulated. Components of any frequency may be obtained when signals of different frequencies are mixed.

It is common practice to use a *diode mixer* in receivers, especially at ultrahigh frequencies (Fig. 11-29). In the diode mixer, the voltages of the frequencies being mixed are fed to the anode circuit. In its lower portion the diode characteristic curve is close to a quadratic parabola and, therefore, the anode current is proportional to the square of applied voltage, i.e.;

$$i_a = k u_a^2 \quad (11-32)$$

where k is a constant.

Under the action of both voltages the resultant voltage will be

$$u_a = V_{a1} \sin 2\pi f_1 t + V_{a2} \sin 2\pi f_2 t$$

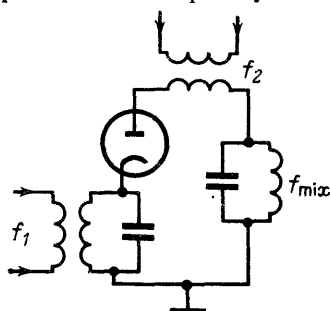


Fig. 11-29. Diode mixer

Assume that both voltages are equal, then according to formula (11-32) the current in the anode circuit will be given by

$$\begin{aligned} i_a &= kV_a^2 (\sin 2\pi f_1 t + \sin 2\pi f_2 t)^2 = \\ &= kV_a^2 (\sin^2 2\pi f_1 t + 2 \sin 2\pi f_1 t \sin 2\pi f_2 t + \sin^2 2\pi f_2 t) \end{aligned}$$

Transforming the second term of this expression, we obtain

$$2 \sin 2\pi f_1 t \sin 2\pi f_2 t = \cos 2\pi (f_1 + f_2) t - \cos 2\pi (f_1 - f_2) t$$

From this it follows that currents of the sum and difference frequencies are present in the anode circuit of the diode. The desired converted frequency may be separated with the aid of the circuit tuned to one of these frequencies, which is connected into the diode circuit.

In Chapter X we described a frequency multiplier, i.e., the device for delivering an output wave whose frequency is a multiple of the input frequency. Let us now consider the solution of the reverse problem, i. e., the problem of obtaining at the output of some device referred to as a *frequency divider* an alternating voltage whose frequency is an integral submultiple of the input frequency. This problem may be solved with the aid of a conventional self-excited oscillator whose anode circuit is provided with a circuit tuned to the input frequency divided by an integral number of times. An external source producing voltage u_s whose frequency is divided is connected into the circuit instead of an anode battery (B power supply) (Fig. 11-30). A bias to the grid is supplied from a separate source E_g (C power supply).

Anode current is absent during the negative half-cycle of the anode voltage. During the positive half-cycle an anode current pulse appears in the anode circuit which initiates oscillations in the circuit tuned to frequency ω/n . Owing to the presence of the feedback path these oscillations produce across the grid an AC voltage u_g varying at the natural frequency

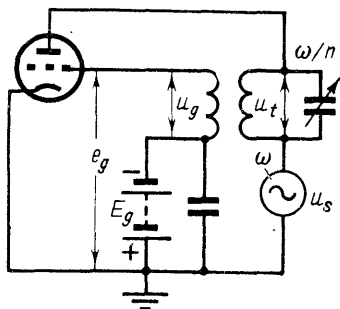


Fig. 11-30. Frequency divider

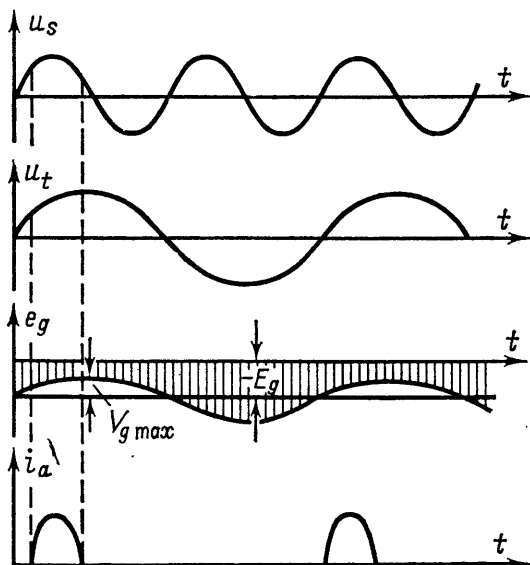


Fig. 11-31. Processes in frequency divider

of this tuned circuit (Fig. 11-31). Because of the negative grid bias, these oscillations cause the current to flow through the valve only when an appreciable positive potential is applied to the anode at the time when the AC component across the grid is also positive.

Assume, for instance, that the external source frequency is twice the natural frequency of the tuned circuit. As can be easily seen from Fig. 11-31, the conditions for the generation of anode current arise once per cycle of oscillations produced in the circuit, anode current pulses replenish its energy and sustained oscillations are produced at a frequency which is twice as less as that of the driving source.

Note that the anode current pulses could pass twice through the circuit for each cycle if there were no feedback and grid bias. One of these pulses would impart energy to the circuit, whereas the other pulse would take it away, so that no natural oscillations could occur in the circuit.

11-7. CONVERSION OF ELECTRIC PULSES

In the preceding chapter we discussed the operating principles of some of the most important pulse oscillators. The pulses obtained with the aid of these oscillators are frequently subject to a great variety of conversions. Therefore, we will discuss now the operating principles of the main converters of electric pulses such as limiters, triggers, differentiating and integrating networks.

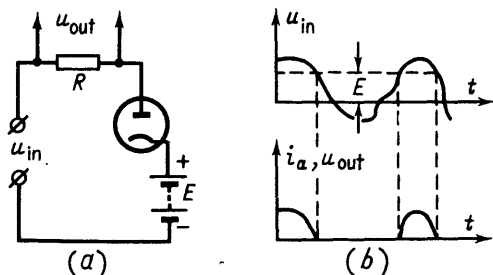


Fig. 11-32. Minimum limiter

Limiters are designed to cut off the signals lying above or below some predetermined level. In other words, current and voltage at the limiter output remain constant if the output voltage exceeds a certain value. If signals are clipped above or below a certain value, the devices are called *maximum* or *minimum limiters* (peak limiters), respectively. Also available are limiters which clip the signal at the top and bottom simultaneously.

The most simple limiter employs a diode which utilizes its property of not passing current when anode voltage becomes less than zero. The circuit of such a limiter (Fig. 11-32a) includes DC source E connected in series with AC source u_{in} . The positive terminal of the DC source is connected to the cathode. Thus, the resultant voltage across the anode is $u_a = u_{in} - E$. Consequently, the diode is conductive only if $u_a > 0$, i.e., $|u_{in}| > |E|$. The resistance of resistor R should considerably exceed the conductive diode resistance equal to hundreds of ohms. When voltage across the anode is negative, no current flows through the diode and no voltage appears

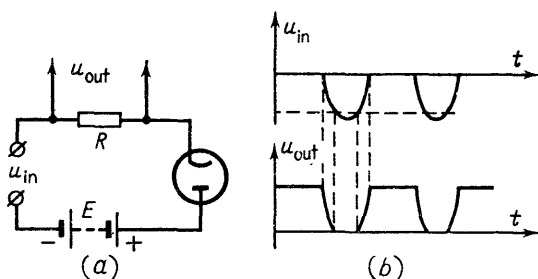


Fig. 11-33. Maximum limiter

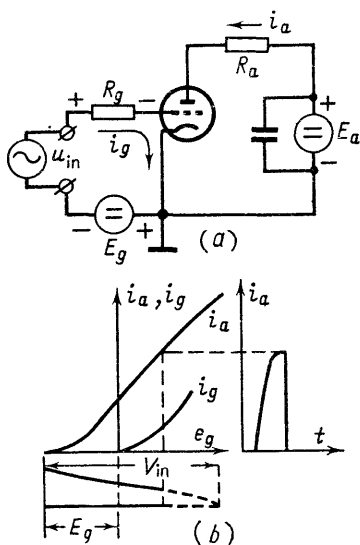


Fig. 11-34. Grid limiter

(a) schematic diagram; (b) processes in grid limiter

across output resistor R . When the anode voltage becomes positive, the diode starts to conduct and a current (proportional to input voltage) will flow through resistor R , thus producing a voltage proportional to its resistance (Fig. 11-32b).

If the reverse problem is posed, i.e., to produce a voltage across the output resistor, which varies proportionally to the input voltage until the latter reaches the value E and then

remains constant, the diode terminals should be interchanged so that negative pulses will be applied to the cathode and the voltage source should have its positive terminal connected to the anode (Fig. 11-33a). Thus, we shall obtain the output voltage as shown in Fig. 11-33b. As compared with the preceding circuit, the pulse produced across resistor R is of the reverse polarity.

The peak limiter (employing the grid-cathode portion of a triode) whose schematic diagram is illustrated in Fig. 11-34a finds extensive application. This circuit differs from the conventional resistance-coupled amplifier by the presence of resistor R_g (having large resistance) in the grid circuit and by the choice of its operating point. Limiting is obtained here when a small positive voltage is applied to the grid. The grid current i_g arising in the circuit as grid voltage becomes positive passes through resistor R_g , thus producing across it the voltage drop which compensates practically a further increase of input voltage u_{in} . Therefore, the grid-cathode voltage remains almost constant and the top of a pulse appearing in the anode circuit proves to be clipped (Fig. 11-34b). Negative swings of input voltage may be also clipped in this circuit by selecting the proper bias voltage E_g .

In some instances it is necessary to have devices capable of producing at their output the current or voltage pulses whose amplitude does not depend on the amplitude, waveform or to some extent upon the duration of the pulses applied to their input. To trigger such a device, the amplitude of the input signal should exceed a certain minimum value. The device of this type represents a peculiar electronic relay which is frequently referred to as a *trigger*.

Figure 11-35a illustrates a typical trigger circuit which resembles the circuit of the multivibrator. It differs from the latter only in the absence of capacitors and the presence of a DC bias source E_g connected in the grid circuit. The trigger reaches its steady state when one of the valves becomes cut off and the other one, conductive.

Assume that equal currents flow at some instant of time through the circuits of both valves. Due to inevitable fluctuations of the electron flow, current in one of the valves will increase somewhat at some instant. Let this occur in valve V1. As this takes place, the voltage drop across the resistor

R_{a1} will become greater, so that a lesser voltage will be applied to resistors R_1 and R_{g2} . As a result, the current flowing through them will decrease and the voltage across the grid of valve $V2$ will drop. This leads to a decrease in the current flowing through the valve and an increase of the voltage across its anode. In turn, the last factor causes an increase in the current flowing through resistors R_2 and R_{g1} and a rise in the grid potential of valve $V1$. The current passing through this

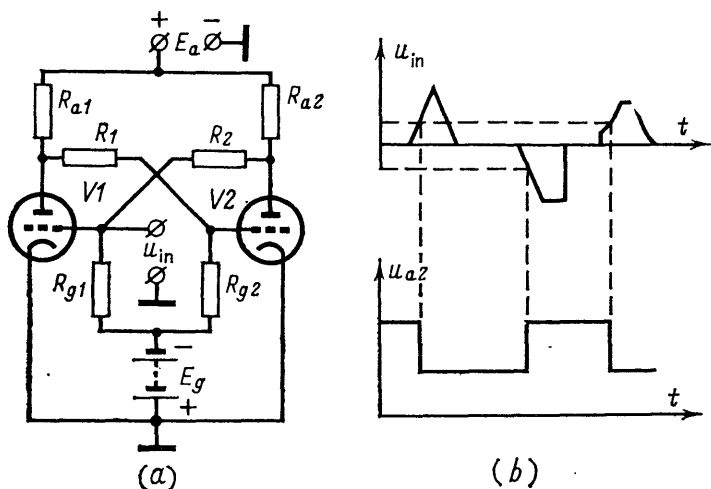


Fig. 11-35. Trigger

(a) schematic diagram; (b) voltages at trigger input and output

valve rises still further. This process results in an avalanche-type decrease of current in one of the valves and a current rise in the other valve until one of the valves becomes cut off. The circuit can remain in this state for an infinitely long period of time, i.e., such a state will be stable.

When, however, the pulse of a positive voltage, sufficient to produce only a small anode current in the cutoff valve, is applied to its grid all the processes will progress in the opposite direction and result in an extremely rapid cutoff of the earlier conductive valve and in the triggering of the earlier cutoff valve. The circuit can be triggered from one state to the other also by applying to the grid of the conductive

valve a negative pulse capable of producing a small decrease in its anode current. Consequently, the circuit can be triggered from one state to the other by feeding the grid of one of the valves with pulses of a proper polarity.

The curves in Fig. 11-35*b* represent anode voltage variations in valve *V*2 when the circuit is triggered under the action of external pulses whose amplitude exceeds the trigger voltage. These pulses are applied to the grid of valve *V*1. When the valve becomes conductive, the pulses control its anode current. The voltage across valve *V*2 changes in constant-amplitude steps.

It is frequently necessary in pulsed devices to convert the signal applied to the input so that the output voltage or current will be proportional to the rate of change of the input signal rather than to its amplitude. Such an operation is called input-signal differentiation, and the circuits in which this process occurs are known as *differentiating circuits*.

A network consisting of a resistor and capacitor connected in series may serve as an example of the simplest differentiating circuit (Fig. 11-36*a*). If the reactance of capacitor *C* for the basic components of pulse current is much greater than the resistance of resistor *R*, the nature of current variations in the circuit is determined primarily by the capacitor. In turn, the current flowing through the capacitor circuit is proportional to the rate of change of voltage across it and is defined as

$$i_C = \frac{\Delta q}{\Delta t} = \frac{C \Delta u}{\Delta t}$$

where $\Delta u / \Delta t$ is the rate of voltage change. While flowing through resistor *R*, this current produces a voltage across it which is proportional to the rate of change of the input voltage.

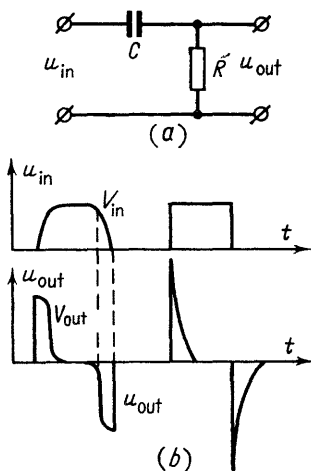


Fig. 11-36. Differentiating circuit

(a) schematic diagram; (b) input and output voltages

Differentiating circuits make it possible to obtain short pulses with a steep leading edge when their input is fed with pulses of almost square waveform (Fig. 11-36*b*). The pulse at the output appears at the instant the pulse is applied to the input. The steeper the leading edge of the input pulse, i.e., the greater the rate of voltage rise, the greater the amplitude of the output pulse. As the rate of the input-voltage rise decreases, the output voltage drops and becomes zero for the time period when the input voltage does not vary (the flat top of the input pulse). When input voltage drops, the voltage pulse appears once again at the output but now with the opposite polarity since the rate of input-voltage change becomes negative. Physically, the different polarity of the two pulses may be explained by the fact that in the first case the capacitor charges and in the second case it discharges.

All this can be easily traced when comparing input and output voltages shown in Fig. 11-36*b* for bell-shaped and square pulses at the input. It can be also seen from this figure that although the amplitudes of both input pulses are equal the output voltage is found to be greater for the square pulse. In the case of an ideal differentiation, the amplitude of the output pulse would be equal to infinity and the pulse would have an infinitely small duration. The presence of the resistor in the circuit, however, results in that the amplitude of output pulses as well as their duration is finite. The less the resistance of the resistor and, consequently, the less the time constant, the closer the pulse waveform to the ideal one.

In some instances, when a current varying in time flows through a circuit, it is necessary to obtain a voltage proportional to the charge which passes through the circuit for some time interval. To determine the amount of this charge, the entire time interval of interest should be broken down into small time fragments within which the current may be considered to be constant with a predetermined degree of accuracy and the sum of the charges transferred during these time fragments should be calculated. This operation is referred to as *integration*.

One may suggest a great variety of circuits which can be used for integration, i.e., produce at their output a voltage proportional to the electrical quantity being integrated (accumulated) which acts at the input. A capacitor is the simplest

integrating circuit. In fact, the voltage across its plates is proportional to the sum of all the accumulated charges which have been carried by the current in its circuit.

Let us consider in detail the integrating properties of the capacitor, C , taking into account that some losses are always present in its circuit (Fig. 11-37a). If a square pulse (Fig. 11-37b) is applied to the input of such a circuit, voltage across the capacitor will rise gradually with its value tending to the amplitude of the input pulse. However, the voltage should not rise during the pulse up to a value comparable with the input voltage otherwise it would prevent current from flowing through the circuit. Therefore, resistor R having a rather large resistance is connected in series with the capacitor, thus making the time constant of the circuit much greater than the pulse duration. On the other hand, the time constant should be small so that on completion of the input pulse the capacitor will have time to discharge before the beginning of the next pulse. When a pulse train has to be integrated, the time constant of the circuit should be increased accordingly.

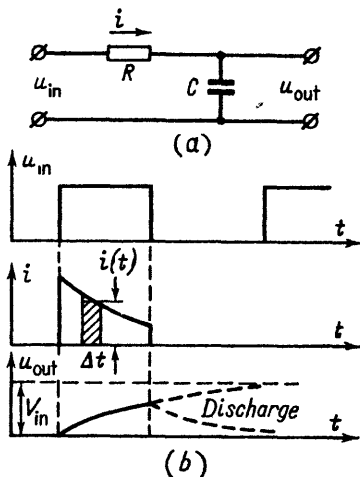


Fig. 11-37. Integrating circuit
(a) schematic diagram; (b) voltages and currents in circuit

In order to deflect an electron beam in oscilloscopes as well as to perform other operations associated with electronic time reading, it is necessary to have devices capable of producing voltages which rise proportionally with time. Naturally, this rise should come to a stop at some instant and then start to rise once again after the desired time interval. In such devices the output voltage has a sawtooth waveform.

A great variety of circuits capable of producing the sawtooth signals can be suggested. We shall discuss here the simplest circuit in which square pulses are converted to the voltage

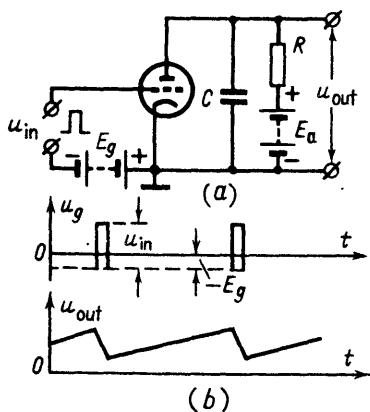


Fig. 11-38. Sawtooth wave generator
(a) schematic diagram; (b) input and output voltages

pulses varying with time by a linear law (Fig. 11-38a). In this circuit, the output voltage is taken from capacitor C connected in parallel to the anode of a valve (triode). When the valve is cut off by voltage from a DC bias source, E_g , the capacitor is charged from an anode voltage source, E_a , through resistor R . The process of voltage rise across the capacitor is shown in Fig. 11-38b. At the initial stage the voltage increases almost linearly. The grid is fed at some instant with a

short positive pulse of the square shape which abruptly triggers the valve. The valve is chosen with its internal resistance as low as possible. Therefore, although the capacitor discharges by an exponential law it does so quickly that the output voltage does not differ essentially from the sawtooth shape. The process is repeated on completion of the trigger pulse.

Chapter XII. RECEIVERS

12-1. PURPOSE AND BASIC CHARACTERISTICS

A receiver intercepts energy of RF signals from the electromagnetic field of incoming waves, converts RF oscillations into primary electric signals and reproduces transmitted intelligence in a visible or audible form. These are compulsory (i.e., basic) functions of any receiver from the simplest to the most complicated one. As radio equipment was complicated, it was entrusted with other functions such as amplification of RF signals, amplification of primary signals (of audio frequency or video pulses), automatic gain control (AGC), etc.

It is now convenient to familiarize the reader with the compulsory functions of receivers by using the simplest circuit which is not used any more but at the dawn of radio communication it was a success. This is the so-called crystal receiver designed for reception of amplitude-modulated telephone signals (Fig. 12-1).

The crystal receiver consists of a receiving antenna A with grounding G and the receiver itself. The input circuit composed of variable capacitor C and inductance coil L is arranged between A and G . These elements are included into the antenna circuit and allow it to be tuned to the desired

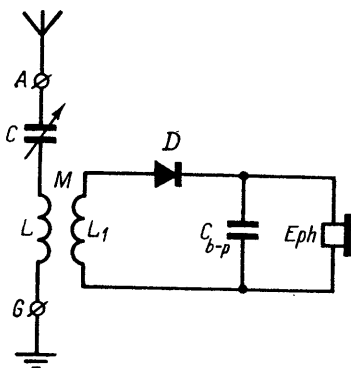


Fig. 12-1. Crystal receiver

frequency, i.e., to the station being received. Compensation of antenna reactance is the condition of receiver tuning, i.e.,

$$X_A + \omega L - \frac{1}{\omega C} = 0$$

It is precisely the resonance that provides selectivity of the simplest receiver, i.e., the ability to separate RF signals of any one station and suppress RF signals from stations operating at other frequencies.

The process of intercepting energy from an electromagnetic field is accomplished in the *antenna circuit*. Incoming waves intersect the antenna wire by the lines of force of their magnetic field inducing in antenna elements (sections of the antenna wire) EMFs whose frequency is equal to the desired signal frequency. The combined action of the elementary EMFs is equivalent to the introduction of a source of EMF having some resulting amplitude into the antenna circuit. Maximum power is developed by this source at resonance in the antenna circuit, i.e., in the absence of reactance. By this means the *incoming wave field energy is converted into the energy of an HF current*. Incoming signals of other wavelengths induce EMFs at other frequencies and, therefore, both resistance and reactance are present in the antenna circuit which prevent them to develop a full HF power. This is the essence of *receiver selectivity*.

Some energy of the useful signal must be *converted into the energy of the primary signal*. For instance, audio (modulating) frequency current serves as the primary signal in the case of amplitude modulation. This conversion occurs in the *detector circuit*. The circuit consists of coil L_1 having coefficient of mutual inductance M with the antenna circuit coil, detector D itself and a primary-signal consumer — earphone E_{ph} which is blocked by the by-pass capacitor C_{b-p} . The coupling between the detector and antenna circuits should provide transmission of the maximum possible signal RF power to the detector.

The detector circuit (coupled with the antenna circuit) is aperiodic. Antenna current induces in coil L_1 an EMF for which capacitor C_{b-p} represents, practically, a short circuit. Therefore, HF voltage taken from the terminals of coil L_1 is fully applied to the detector. The detector constitutes

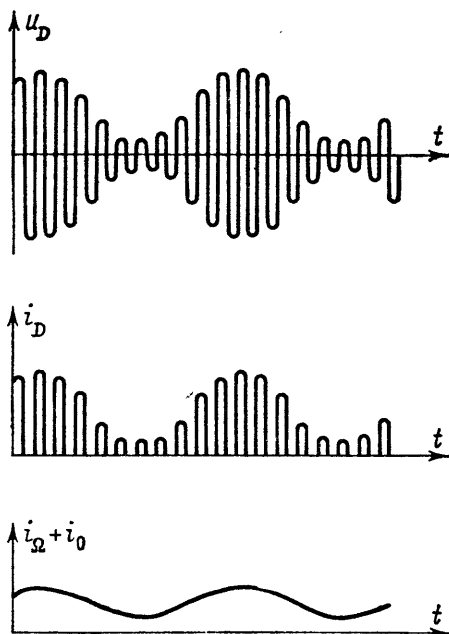


Fig. 12-2. Physical processes in detector circuit

a non-linear device, for instance, a crystal (or vacuum) diode. Under the action of alternating voltage u_D the detector conducts current i_D in one direction only (Fig. 12-2), i.e., a DC component appears in this rectified current, but if the signal is amplitude-modulated, DC component i_0 varies itself, thus inducing an RF-signal envelope, i.e., it contains audio frequency current i_Ω . The earphone which converts the electric signal into audible one serves as a consumer of audio frequency current energy, i.e., as a load.

The simplest receiver makes use of incoming wave energy only and does not have its own supply source. Of course, power sufficient for normal operation of a loudspeaker cannot be obtained since the EMF induced in the antenna is small. One cannot also expect satisfactory interference rejection in such a receiver which has one tuned circuit only. But even this simple example allows us to define the parameters which

characterize the receiver performance and contemplate the ways of improving these parameters.

Sensitivity, i.e., the ability of receiving weak signals, is one of the major properties of a receiver. It is characterized quantitatively by that minimum power or minimum EMF induced in the antenna which provides normal signal reproduction. The less the numerical value of sensitivity, the greater the sensitivity. The simplest (crystal) receiver (Fig. 12-1) has sensitivity which is characterized by an RF signal power of about 10 mW and EMF of at least tenths fractions of a volt, whereas a commercial broadcast receiver has the sensitivity of about 50 μ V. Sensitivity of modern special receivers amounts to hundredth fractions of a microvolt or billionth fractions of a milliwatt.

Sensitivity may be increased by connecting *amplification stages* into the receiver both for an HF channel (predetection stage) and LF channel (post detection stage). Certainly, one should try to obtain the sensitivity which is actually necessary for a given receiver since signal distortions and instability of receiver performance increase with the increasing number of amplification stages. Probability of signal suppression by extrinsic and intrinsic interference (external and internal noise) also grows with an increase of total gain. Five or six high-, intermediate- and low-frequency amplification stages are generally used in valve broadcast receivers of medium-grade. In transistor receivers which are more economic from the power supply viewpoint the number of amplification stages may be the same or somewhat more.

The second factor of the quality of a receiver is its frequency *selectivity*, i.e., the ability of the receiver to ensure normal reception of the signal of the desired frequency while rejecting signals of all other frequencies. In other words, selectivity may characterize the degree of protection of the receiver against the action of undesirable transmitters which prove to be the most probable source of interference in medium- and short-wave bands.

Frequency selectivity is achieved by the use of *tuned circuits* in the receiver stages. Selectivity of a receiver as a whole may be characterized approximately by the shape of its resonance characteristic curve, i.e., by the relationship between selectivity and the difference between the transmit-

ter frequency and the receiver operating frequency. Fig. 12-3 illustrates an ideal and actual resonance characteristic curves of a receiver. The frequency difference (in kilohertz) is given along the X axis, and the Y axis represents the ratio of output voltages of non-resonant signals V to the resonant-signal voltage V_0 for equal EMF in the antenna. The ideal shape of the resonance characteristic curve from the viewpoint of selectivity would be rectangular since it would provide identical conditions for the propagation of frequencies which are within the signal band and complete suppression of all other frequencies.

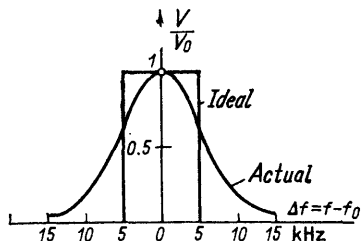


Fig. 12-3. Resonance characteristic curves of receiver

In practice, however, such an ideal resonance characteristic cannot be obtained but it may be approximated by using tuned circuits in RF amplifier stages. The closer the shape of

an actual resonance characteristic to the ideal one, the higher the receiver selectivity. Consequently, apart from the function of amplification of desired signals, RF stages are responsible for increasing the receiver selectivity.

For the simplest receiver (Fig. 12-1) in which no amplifier stages are present, selectivity is determined by the antenna circuit only and, therefore, it proves to be very poor.

When in determining the notions of "sensitivity" and "selectivity" we talked about the "normal" effect at the receiver output, we actually implied the *output power* which is *normal* for a given receiver (rated output power). For a given loudspeaker, we may talk about the normal value of voltage across its terminals. The maximum value of broadcast receiver output power is limited by non-linear distortions which arise whenever the output stage is overloaded (see Chapter IX).

Rated output power should suit the sizes of the room where the receiver is intended to operate and the level of external noise. The power required for a headphone is measured by milliwatts only. Domestic broadcast and TV receivers have the rated output power from tenths of a watt to several watts.

The band of frequencies (waves) at which the receiver may be tuned is a very important quality index of its properties and capabilities. Characteristic of broadcast receivers is the division of their total band into subbands. Subbands are switched over by changing the coils in HF circuitry. Continuous tuning within each subband is accomplished with the aid of a variable capacitor unit. For instance, in the simplest receiver (Fig. 12-1) the subband may be switched over by

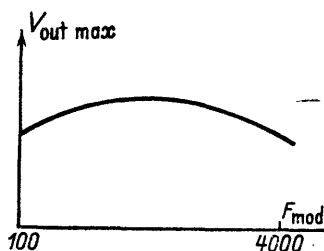


Fig. 12-4. Over-all frequency response of receiver

changing the coil L together with which is changed the coupling coil L_1 . Subbands are numbered with figures which increase with the increasing frequency.

Modern broadcast receivers have subbands of long, medium, short and ultrashort waves. Instead of a single (continuous) short-wave band they usually have several narrow short-wave subbands with "extended" scales

for the bands which are overloaded with broadcast stations.

A broadcast receiver is also characterized from the viewpoint of *fidelity* in the reproduction of a primary signal (speech or music). Reproduction quality depends first of all on uniformity of amplification of signals (modulating signals) extracted from the RF received signals. High-, intermediate- and low-frequency amplifier stages may be the cause of non-uniformity in the passage of a primary signal spectrum through the receiving channel. If the stage bandwidths of tuned or bandpass amplifiers are chosen too narrow, in such stages the side frequencies produced by higher modulating audio frequencies, which restore audio frequency in the process of detection are amplified insufficiently. If the frequency response of low-frequency stages is too non-uniform, the signals of both upper and lower audio frequencies may be amplified insufficiently.

Fidelity of a primary signal is estimated from the overall frequency response of the receiver which represents the relationship between the output voltage (across the loudspeaker) and the modulation frequency (at a fixed EMF induced in the antenna and a fixed modulation depth). An example of the

receiver frequency response is given in Fig. 12-4. When designing a receiver, the permissible irregularity of its frequency response is specified similarly to that of AF amplifiers (see Chapter IX).

Receivers may be classified under several types by a number of indications. On the basis of a circuit configuration they may be subdivided into *direct-detection receivers* and *superheterodyne receivers*.

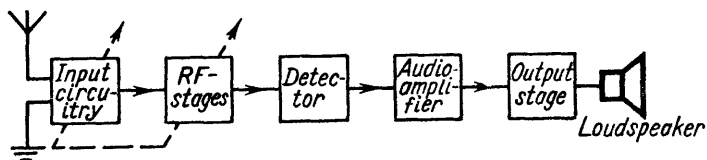


Fig. 12-5. Block diagram of direct-detection receiver

In a direct-detection receiver the frequency of the radio signal being received is converted to that of the primary signal by means of detection. The block-diagram of such a receiver (Fig. 12-5) consists of an input circuitry connected to the antenna, radio frequency amplifier stages, a detector and an audio-frequency amplifier terminating in an output stage. Subjected to tuning are the input circuits and the RF amplifier stages (or more precisely the tuned circuits incorporated in these units). In this case tuning is of a ganged type accomplished with the aid of a ganged capacitor.

The direct-detection receiver is identical in physical processes with the crystal receiver differing only by the fact that the latter is added with two amplifiers operating in two frequency bands. The deficiencies inherent in the crystal receiver are not eliminated to a large extent in the direct-detection receiver. For instance, its passband in the short-wave range will be always wider than that required in radio-telephony. In addition, its sensitivity changes significantly over the passband due to variations in circuit parameters. At present direct-detection receivers are not manufactured any more. In some instances, radio amateurs start their practical activities from assembling transistor direct-detection receivers.

In the superheterodyne receiver frequency is converted at least twice. Frequency of the radio signal being received is converted to an *intermediate frequency (IF)* which is constant for a given receiver. In the process of detection the intermediate frequency is converted to the audio frequency of the primary signal. In conformity with this the block-diagram of a superheterodyne (Fig. 12-6) consists of an input circuitry, RF amplifier stages, frequency converter, IF amplifier stages, detector, and an audio-frequency amplifier with an output stage. Subjected to tuning are the input

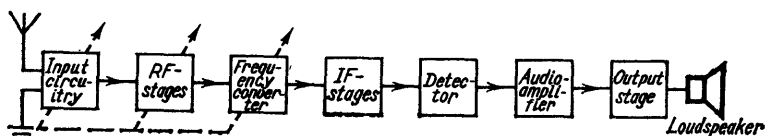


Fig. 12-6. Block diagram of superheterodyne receiver

circuits, the high-frequency amplifier and the heterodyne of the frequency converter. In this case, tuning should be of the ganged type. The name "superheterodyne" which has appeared since the dawn of valve radio engineering should imply that the frequency obtained as a result of the first conversion was a lowered radio frequency rather than an audio frequency.

Fixed tuning of IF amplifier stages makes it possible to use improved filters providing high selectivity. In addition, amplification in three different frequency regions allows high sensitivity to be obtained without the risk of receiver self-excitation. These properties made the superheterodyne essentially the main circuit configuration used in radio systems. Two or even three frequency converters may be used in special superheterodyne receivers. Broadcast receivers, however, are generally built in accordance with the block-diagram shown in Fig. 12-6.

As to the type of active devices used, receivers may be classified under two types, *valve* and *transistor receivers*. Until recently, the valve receiver was considered to be capable of displaying higher fidelity than that of the transistor receiver. Progress, however, made in semiconductor technology is at variance with this conclusion. Moreover, transistor

equipment is smaller in size and more economic from the viewpoint of power consumption.

By their properties and design, commercial broadcast receivers are subdivided into three classes. The first class receiver displays the best performance, though its circuitry and design are most complicated. More simple and inexpensive receivers fall in the second and third classes.

12-2. RECEIVER INPUT CIRCUITS

A receiver (see Figs. 12-5 and 12-6) is composed of functional units which we have already discussed in Chapter IX. Among these units are AF amplifier stages. The remaining units (input circuits, RF and IF amplifiers, frequency converter, detector), however, deserve special consideration as well. Furthermore, a receiver may include some auxiliary devices improving receiver performance (for instance, automatic and manual controls, visual tuning indicators, etc.) which also require consideration since only after this may we discuss receiver circuits in their entirety. We begin the study of functional units with receiver input circuits.

The receiver *input circuitry* serves as an intermediate component between the receiving antenna and the first amplifier stage. An example of the input circuit we have seen in Fig. 12-1 where the antenna circuit included some elements used for tuning the receiver to the signal frequency. The receiver, however, should operate from antennas of different size and parameters (outdoor, indoor and automobile antennas, etc.). It is natural that the tuning of any antenna ganged with that of tuned amplifier stages is not feasible. Therefore, the antenna circuitry is not retuned. It generally incorporates an inductor (sometimes an inductor and capacitor) establishing in the antenna circuit the frequency which lies beyond the subband being received. The untuned antenna is coupled with a closed tuned circuit which is the main part of the receiver input circuitry and is tuned to the frequency of the signal being received simultaneously with tuned-amplifier stages (ganged tuning). The RF signal taken from the terminals of this tuned circuit is applied to the input of the 1st amplifier stage. When a subband is changed, the tuned-circuit inductor is switched over simultaneously with the antenna-circuit

inductor. If a set of trimming capacitors is connected into the circuit, they are switched over as well.

The antenna circuit may be coupled either directly or inductively with the closed tuned circuit. A schematic diagram of an inductive-coupled input circuit is shown in Fig. 12-7. In the case of a valve receiver whose input stage operates practically with no losses in the control

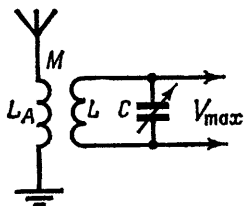


Fig. 12-7. Inductive coupling between antenna and tuned circuit

grid circuit, the complete signal voltage $V_{\max}$ is taken from the terminals of the capacitor C . Input circuit efficiency is estimated with the aid of the so-called *transmission gain* K . The transmission gain is the ratio of the voltage $V_{\max}$ taken from the circuit to the EMF $E_{\max}$ induced in the antenna by the signal being received, i.e.,

$$K = \frac{V_{\max}}{E_{\max}} \quad (12-1)$$

The EMF induced by the signal in the simplest receiving antenna equals the product of the electric field intensity $E_{\max}$ of incoming waves and the antenna effective height h_{eff} , i.e.,

$$E_{\max} = E_{\max} h_{\text{eff}} \quad (12-2)$$

For instance, if wave field intensity $E_{\max} = 10 \mu\text{V/m}$ and indoor antenna effective height $h_{\text{eff}} = 2\text{m}$, then electromotive force $E_{\max} = 100 \times 2 = 200 \mu\text{V}$. If the voltage $V_{\max} = 1\text{mV} = 1000 \mu\text{V}$ is taken from the tuned circuit terminals, the voltage transmission gain $V_{\max}/E_{\max} = 1000/200 = 5$.

The voltage transmission gain is determined by circuit parameters. In the case of an untuned inductive-coupled antenna the transmission gain may be determined by using the theory of coupled circuits from Chapter III. Let us assume that at broadcast-band frequencies the input impedance of the antenna itself obtained from formula (6-8) is a capacitive one at the signal frequency ω

$$X_A = \frac{1}{\omega C_A}$$

Antenna circuit reactance will be $\omega L_A - 1/\omega C_A$, where L_A is the inductance connected to the antenna and coupling it to the closed LC -circuit.

The EMF induced by a signal in the closed circuit, $E_{t \max}$, may be obtained if we neglect the antenna circuit resistance as compared with its reactance and express the current in the antenna by dividing the antenna EMF, $E_{\max}$, by the antenna impedance, thus we may write

$$E_{t \max} = I_{A \max} \omega M \approx \frac{E_{\max} \omega M}{\omega L_A - \frac{1}{\omega C_A}} = E_{\max} \frac{M}{L_A \left(1 - \frac{1}{\omega^2 L_A C_A}\right)}$$

As is known from the theory of resonance (see formula 2-46), the voltage taken from the closed circuit capacitor at series resonance, $V_{o \max}$, proves to be Q_{eq} times greater than the EMF in the same circuit, i.e., $V_{o \max} = E_{t \max} Q_{eq}$. Here, Q_{eq} is the Q -factor of the tuned circuit computed with allowance made for losses inserted into the circuit due to coupling to the antenna and due to the connection to the 1st amplifier stage. By referring to the value of $1/\sqrt{L_A C_A}$ as the antenna circuit natural frequency ω_A , we may find the resonance voltage transmission gain of the input circuit

$$K_0 = \frac{V_{o \max}}{E_{\max}} = \frac{M}{L_A \left(1 - \frac{\omega_A^2}{\omega_0^2}\right)} Q_{eq} \quad (12-3)$$

where $\omega_0 = 1/\sqrt{LC}$ is the closed circuit natural frequency equal to the signal frequency when the circuit is tuned to resonance.

Mutual inductance M is chosen to be rather small and transmission gain does not generally exceed 10 so that antenna coupling could not affect too much the closed circuit parameters. It should be borne in mind that voltage is increased (transformed) here at the expense of resonance but there is no power amplification since the input circuit contains no active elements.

The resonance characteristic of the input circuit (see Chapter II) is a curve which agrees with equation (2-50) and is expressed by the formula

$$y = \frac{V_{\max}}{V_{o \max}} \approx \frac{1}{\sqrt{1 + Q_{eq}^2 \left(\frac{2\Delta\omega}{\omega_0}\right)^2}} = \frac{1}{\sqrt{1 + Q_{eq}^2 \left(\frac{2\Delta f}{f_0}\right)^2}} \quad (12-4)$$

Here, $\Delta f = f - f_0$ is the *detuning factor*, i.e., the difference between the signal and circuit resonant frequencies. To estimate selectivity of the input circuit, it is necessary to compare its resonance characteristic curve with an ideal one (see Fig. 12-3). *Bandwidth* B , i.e., the frequency band within which the signal spectrum is contained (see Fig. 11-4) may be determined from formula (2-51)

$$2\Delta f_{0.7} = B = \frac{f_0}{Q_{eq}} \quad (12-5)$$

For instance, if the Q -factor of a tuned circuit equals 100, its bandwidth will amount to 1 per cent of the carrier frequency. The bandwidth is limited by the ordinates making up $1/\sqrt{2} \approx 0.7$ of the resonance ordinate.

The ideal shape of a resonance characteristic provides complete suppression of interference whose frequencies lie beyond the passband of the circuit. In other words, the interference band equals the passband of the circuit. In practice, an actual circuit is subjected also to interference from transmitters outside of its passband. The portion of the circuit frequency response between ordinates comprising 0.1 of the resonance ordinate may be considered as a *hazardous interference band*. Assuming that $y=0.1$ in formula (12-4) and solving it with respect to $2\Delta f$, we obtain:

$$2\Delta f_{0.1} \approx \frac{10f_0}{Q_{eq}}$$

The passband-to-interference band ratio is known as the *form factor* of a resonance characteristic. It is evident that for an ideal characteristic this factor equals unity (rectangularity 100 per cent). For a single tuned circuit we have

$$\frac{B}{2\Delta f_{0.1}} = \frac{2\Delta f_{0.7}}{2\Delta f_{0.1}} \approx 0.1 \quad (12-6)$$

i.e., the form factor amounts to only 10 per cent. This is a poor selectivity leaving high probability for interference effect. To improve the form factor of the over-all resonance characteristic curve of a receiver, one tuned circuit may be connected into each RF amplifier stage. If the total number of identical circuits is N , the over-all resonance characteristic

curve y_{0-a} will be described by the equation

$$y_{0-a} = y^N = \frac{1}{\left[\sqrt{1 + Q_{eq}^2 \left(\frac{2\Delta f}{f_0} \right)^2} \right]^N} \quad (12-7)$$

It is quite evident that ordinates y 's which are close to unity change little when being raised to the power, whereas small fractions decrease abruptly. This means that the form factor in this case will be greater. It can be shown that when the number of circuits is increased indefinitely, the form factor will increase tending to a limit equal to 39%. This is also not a good way of improving receiver selectivity. It would be much more expeditious to insert bandpass filters into the amplifier stages. Practically, bandpass filters are employed in IF amplifier stages having a fixed tuning.

Thus, we have gained some idea of the estimation of input circuit sensitivity (voltage transmission gain) and selectivity (the form factor of the frequency response). A *transmission gain non-uniformity within some frequency band* (i.e., when the circuit is retuned) is the third most important parameter characterizing the performance of an input circuitry. In this respect an inductive-coupled circuit may be estimated by formula (12-3). The values of M and L_A do not change as the circuit is retuned. When the circuit is retuned with the aid of a variable capacitor, the Q -factor (Q_{eq}) may also be considered to be constant ($Q = \omega_0 L / r$) since the resistance r characterizing the losses in the circuit increases with the increasing frequency ω_0 . Consequently, the transmission gain non-uniformity will be determined in formula (12-3) by the ratio ω_A / ω_0 . It is advantageous to have $\omega_A < \omega_{0 \min}$, i.e., to tune the antenna circuit to the frequency lying *below the lowest band frequency* by an appropriate choice of the inductor L_A . In fact, if $(\omega_A / \omega_0)^2 < 1$, frequency adjustment, i.e., variations in ω_0 will affect little the value of K_0 . The transmission gain non-uniformity within the circuit frequency band will be insignificant. As this takes place, the transmission gain may somewhat decrease with the increasing frequency ω_0 .

As far as the input circuit bandwidth B is concerned, it extends proportionally to the frequency ω_0 . This is evident from formula (12-5). The band performance of inductive-coupled input circuits when $\omega_A < \omega_{0 \min}$ is illustrated in

Fig. 12-8 which also arbitrary shows absolute values of transmission gain for three operating points within the receiver pass-band. The decay of transmission gain in the higher frequency region is very useful in those instances when the succeeding amplifier stages increase their gain with the increasing frequency.

Figure 12-9 illustrates the same inductive-coupled circuit of an untuned antenna and tuned circuit but in this case it is designed for two bands of long and medium waves and completed with an appropriate ganged switch and trim-

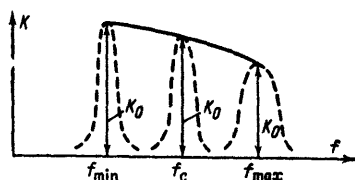


Fig. 12-8. Schematic representation of band properties of inductive-coupled input circuit

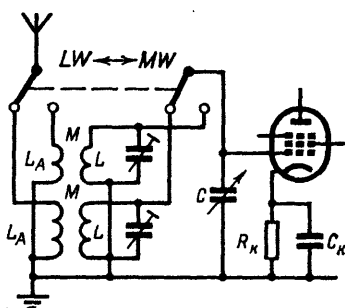


Fig. 12-9. Inductive-coupled input of two-band valve receiver

mer capacitors. A bias voltage is applied to the valve grid from the lower end of resistor R_k through inductor L .

The connection of an antenna to an input circuit through a low-capacitance blocking capacitor C_A (Fig. 12-10a) is a simpler version of an input circuit used in valve receivers. The connection of the small capacitance C_A weakens sharply the effect of the antenna on the circuit tuning and does not disturb the gang of its tuning with that of other stages accomplished with the aid of a gang capacitor. The voltage transmission gain decreases with the decreasing capacitance C_A and generally it also does not exceed 10.

It is different with the design of an input circuit used in transistor receivers where the input impedance of the 1st stage is low as compared with that of a valve. So that the shunting action of the transistor input circuit cannot reduce too abruptly the equivalent Q -factor, Q_{eq} , of the LC tuned circuit, coupling to the transistor base circuit is accomplished through

a step-down autotransformer (Fig. 12-10b) or a conventional transformer.

In Fig. 12-10b between the outdoor antenna A and the LC circuit is shown the capacitive coupling. As is seen from the figure, however, there is a possibility of switching off the outdoor antenna and using the inductor L as a *magnetic antenna* (MA).

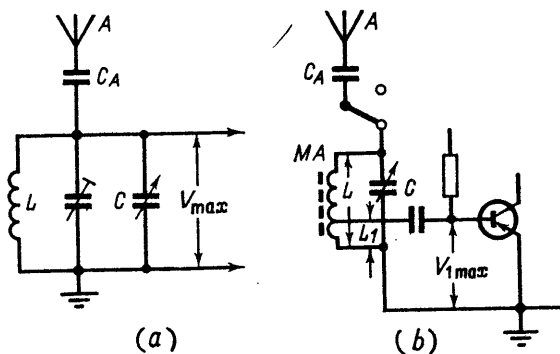


Fig. 12-10. Versions of input circuits

(a) input circuit capacitively coupled with antenna; (b) "magnetic" and outdoor antennas of transistor receiver

The magnetic (ferrite) antenna which is convenient in use since it is contained inside the receiver presents a variety of the loop antennas (see Chapter VI). The magnetic antenna consists essentially of a coil wound on a ferrite core. Owing to the high permeability of the core, the magnetic field of incoming waves sets up an intensive magnetic flux in the coil and induces in it a noticeable EMF even when its sizes are very small. Such an antenna possesses directive reception similarly to the loop one, i.e., a maximum received signal is obtained when the ferrite core is oriented along the magnetic lines of force (perpendicular to the beam) of incoming waves. Naturally, the magnetic antenna should be placed outside of the screen used for shielding the receiver's front end. Directional properties of the magnetic antenna are realized either by rotating the antenna inside the receiver or by turning the receiver itself.

The connection of the transistor to the magnetic antenna circuit is estimated quantitatively by the transformation ratio $m = V_{1\max}/V_{\max} \approx L_1/L$, where L is the total circuit inductance; $V_{1\max}$, the output voltage; $V_{\max}$, the voltage across inductor L ; and L_1 , the inductance of that portion of coil turns to which the transistor is connected. In the case of inductive coupling $m \approx M/L$, where M is the coefficient of mutual inductance between the coupling coil and circuit coil. The greater the value of M , the greater the additional resistance inserted into the antenna circuit by the transistor input. In other words, the Q -factor of the antenna circuit is reduced.

At the band highest frequency the Q -factor is allowed to reduce not more than twice. One may accept $Q_{eq} \approx 0.7Q_c$, where Q_c is the unloaded quality factor of a circuit. It will be recalled that from formula (2-69) the circuit conductance presented by the parallel branch at resonance will be

$$G_c = \frac{1}{r_{eq,r}} = \frac{r_t}{\rho^2} = \frac{1}{Q_c \rho} \quad (12-8)$$

By using transformation principles, transistor input conductance may also be reduced (referred) to the circuit in the form of an additional parallel branch ΔG :

$$\Delta G = m^2 G_{in}$$

where G_{in} is the input conductance of the transistor stage at the signal frequency. The total conductance determines the equivalent quality factor of the circuit

$$G_c + m^2 G_{in} = \frac{1}{Q_{eq} \rho} \approx \frac{1}{0.7 Q_c \rho} \quad (12-9)$$

The transformation ratio m may be found from the above equation.

Assume, for instance, that at the signal highest frequency $\rho = \omega L = 300 \Omega$ and the circuit quality factor $Q_c = 33$. Then the intrinsic parallel conductance of the antenna circuit will be $G_c = 1/Q_c \rho = 1/33 \times 300 \approx 10^{-4}$ S. Let us now assume that the transistor input conductance $G_{in} = 10^{-3}$ S. By substituting these values into formula (12-9), we obtain

$$10^{-4} + m^2 10^{-3} \approx \frac{1}{0.7 \times 33 \times 300}$$

hence

$$m \approx \frac{L_1}{L} \approx \frac{\omega_1}{\omega} = \sqrt{0.04} = 0.2$$

i.e., the position of the tap along the antenna coil must correspond approximately to the ratio $\omega_1/\omega=1/5$.

In case of a magnetic antenna the voltage transmission gain will be

$$K_0 = \frac{V_{1\max}}{Eh_{eff}} = Q_{eq} m$$

In our example $K_0=0.7 \times 33 \times 0.2 \approx 4.6$. This shows that voltage transmission gain is reduced rather appreciably at the transistor input as compared with that of the circuit itself which would be $K_0=Q_c=33$.

In ultrashort-wave receivers even the 1st valve stage exhibits a comparatively low input impedance due to losses in the glass envelope, electrodes and leads. Therefore, the valve is autotransformer-coupled to the tuned circuit. An antenna coaxial feeder is also autotransformer-coupled to the circuit (Fig. 12-11). The tuned circuit functions here not only as a selective tunable system but also as a transformer which matches the feeder impedance to the valve input impedance. As in the case of a transistor-receiver input circuit, it is appropriate to deal here with the power transmission gain from the antenna to the 1st stage input. This gain is always less than unity since the input circuit does not contain active elements.

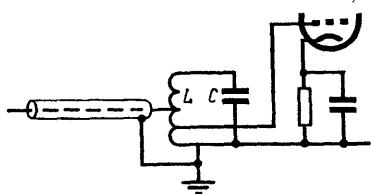


Fig. 12-11. Input circuit of ultra-short-wave receiver

12-3. HIGH-FREQUENCY AMPLIFIERS

As noted earlier, HF amplifier stages are connected between the antenna and the detector. Their mission is to boost radio signal power (voltage or current) up to the value which is high enough to be successfully converted by the detector into a primary signal. It might be well to point out that in

practice a voltage of the order of dozens and hundreds of microvolts is taken from the antenna circuit terminals to the 1st stage input, whereas voltage of about one or even several volts is required for operation of the detector. Consequently, HF voltage must be amplified dozens of thousands of times (whereas power must be amplified many millions of times).

It is common practice in modern broadcast receivers (see Fig. 12-6) to distribute this great amplification between HF and IF amplifiers. HF amplifiers should be tuned each time to the frequency of one or another radio station being received.

After amplification, any radio signal is converted from high to intermediate frequency which is quite definite for a given receiver and falls also within the RF band. The signal is then amplified at this intermediate frequency. In principle, IF amplifier stages are identical with HF amplifier stages but they do not require retuning and, therefore, may be much simpler in design and have higher performance. We shall consider IF amplifiers as a separate kind of amplifiers. It should be emphasized that the AF amplifiers discussed somewhere earlier may be used independently, whereas HF and IF amplifiers are integral parts of a radio receiver only.

A radio signal is an HF oscillation modulated by a control or primary signal. In the presence of modulation the signal no longer remains "monochromatic", i.e., the one containing a single frequency. Modulation introduces additional components which in the simplest case are located on the frequency scale to the right and left of the fundamental oscillation, thus forming a radio-signal spectrum (Fig. 12-12). In order to receive such a signal, it is advantageous to utilize an amplifier possessing selectivity properties, i.e., capable of arranging the radio-signal spectrum within its passband and suppressing the interference at frequencies outside of this passband.

Selectivity properties are attained due to the fact that a band tuned circuit is used as a load in the output circuit of an HF amplifier stage. As a rule, the circuit is tuned with the aid of a variable capacitor and only seldom an inductor or a combination of an inductor and capacitor (adjusted simultaneously) is used for this purpose.

In valve receivers HF tuned amplifiers generally employ pentodes. Special triodes are used only in the band of ultra-

short waves. Pentodes are favourable because of their high anode (internal) resistance, small anode-control grid capacitance (see Chapter VIII), etc. A schematic diagram of a tuned stage employing a pentode is shown in Fig. 12-13a together with the input of the succeeding stage.

Components incorporated in the stage circuit serve the following purpose. Bias voltage is supplied to the control grid from resistor R_k (hundreds of ohms) which is shunted by capa-

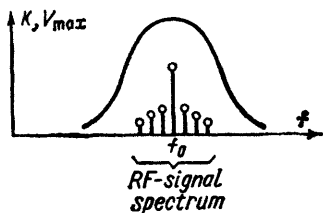


Fig. 12-12. Radio signals within passband of HF amplifier

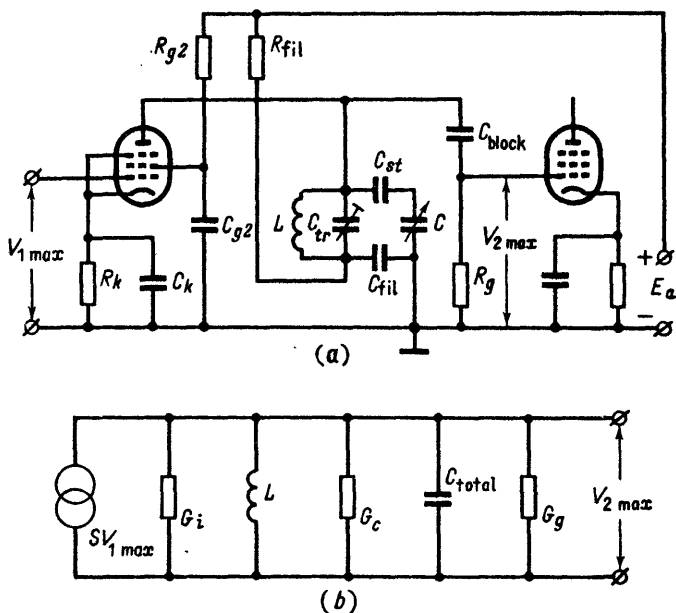


Fig. 12-13. Tuned amplifier stage with direct (full) connection of tuned circuit to anode circuit

(a) schematic diagram; (b) equivalent circuit

citor C_k (about $0.1 \mu F$). It should be noted that operation of a valve in the linear portion of its characteristic curve wit-

hout control grid current can be much simpler provided in a tuned amplifier than in the output stage of an LF amplifier since in the former the amplitudes of the signals being amplified are comparatively small. The screen grid is fed through voltage-dropping resistor R_{g2} (of the order of hundreds of kilohms) being by-passed for high frequency by capacitor C_{g2} (about 0.01 to 0.1 μ F).

Between the anode and cathode of the valve a parallel circuit (see Chapter II) is formed by inductor L (which may have a magnetic core) and variable capacitor C (a part of a gang capacitor). The values of L and C are determined by the frequency band of the receiver. Several subbands are generally required to cover one frequency band (subbands are changed over by switching coils L having different inductances).

To eliminate the effect of the capacitance formed between the operator's hand and the capacitor plates on circuit tuning the capacitor rotor must be connected to the metal chassis of the receiver. However, in case of accidental closure of the stator plates against the rotor plates, a short circuit may occur between the positive terminal of the anode supply source and the negative terminal through the inductor L . To avoid this, a high-capacitance (dozens of thousands of picofarads) stopping capacitor C_{st} is connected in series into the circuit. Feedback between the preceding and succeeding stages through the internal resistance of the supply source is eliminated by the familiar decoupling filters composed of capacitor C_{fil} (dozens of thousands of picofarads) and resistor R_{fil} (several kilohms). Capacitances of capacitors C_{st} and C_{fil} are so high as compared with tuned circuit capacitance C that their series connection affects little the tuned circuit frequency. It is common practice to connect a trimmer capacitor C_{tr} in parallel with coil L , which are both replaced if necessary. Strictly speaking, the resultant circuit capacitance which determines the resonant frequency equals

$$C_{total} = \frac{CC_{fil}C_{st}}{CC_{fil} + CC_{st} + C_{fil}C_{st}} + C_{tr} + C_c$$

where the value of C_c includes the coil capacitance, the output capacitance of the given valve, the input capacitance of the

succeeding stage, and capacitance between the connecting wires.

The HF voltage produced across the tuned circuit by a radio signal is fed to the grid of the succeeding valve through blocking capacitor C_{block} having a capacitance of the order of hundreds of picofarads and good insulation between the plates. To apply negative bias to the control grid of the succeeding valve, it is connected to the cathode resistor through grid resistor R_g (hundreds of kilohms). The components listed above are the main parts of a tuned stage employing a pentode and direct connection of the tuned circuit to the anode circuit.

Let us examine this stage and determine its gain bandwidth and variations of its parameters within the tuning range. For this purpose, we shall use the equivalent circuit shown in Fig. 12-13*b*. C_{total} is the resultant capacitance indicated above; $G_i = 1/R_i$ is the output conductance of the pentode; $G_c = 1/Q_c\rho$, the parallel conductance of the tuned circuit; L , the tuned circuit inductance; $G_g = 1/R_g$, the grid-leak conductance; $SV_{1\max}$, the signal current source in the anode circuit; $V_{2\max}$, the amplified voltage. At radio frequencies the low reactance of capacitor C_{block} may be neglected considering it to be a short circuit for alternating current.

When the circuit is tuned to the signal frequency, the reactances of the inductive and capacitive branches cancel each other (i.e., parallel resonance occurs, see Chapter II), so that the current source will be loaded with the total conductance only and will produce across it an output voltage

$$V_{2\max} = \frac{SV_{1\max}}{G_i + G_c + G_g} = \frac{SV_{1\max}}{G_{total}}$$

At resonance the voltage gain for the stage is

$$K_0 = \frac{V_{2\max}}{V_{1\max}} = \frac{S}{G_{total}} \quad (12-10)$$

Assume, for instance, that $S = 10$ mS; $G_{total} \approx G_c = 0.25$ mS. Then,

$$K_0 = \frac{S}{G_{total}} = \frac{10}{0.25} = 40$$

A susceptance (either capacitive or inductive) appears in the circuit as it goes off resonance, its admittance increases and the gain falls off. It is precisely this relationship that has been shown in Fig. 12-12. We shall call it the resonance characteristic of the amplifier stage.

What determines an attainable gain K_0 ? When a valve is chosen, the circuit quality factor, Q_c , which is reciprocal of its resonance conductance ($Q_c = 1/G_c\rho$) plays a dominant role. Gain may be improved by selecting the *tuned circuit with high Q*. This way, however, has some limitations. First, it is difficult to realize a closed circuit with quality factor Q_c higher than 100 if coil L is home-made. Second, when the conductance G_c is low, conductances G_i and G_g will have a shunting effect on the circuit, thus limiting the gain. Third, the bandwidth may prove to be too narrow when the circuit quality factor is high at long waves (about which we shall talk later). Finally, stage operation may be unstable, i.e., self-excitation may occur in the circuit due to the presence of feedback through the anode-grid capacitance C_{ag} . V. I. Siforov has shown that the stable gain of a tuned stage is limited by the value

$$K_{0\max} \approx 0.5 \sqrt{\frac{S}{\omega C_{ag}}}$$

For instance, if we use a pentode having transconductance $S = 5 \times 10^{-3} \text{ S}$, $C_{ag} = 0.01 \text{ pF} = 1 \times 10^{-14} \text{ F}$, and if $\omega = 2\pi f = 5 \times 10^7$, then $K_{0\max} = 50$. For broadcast bands, the gain ranges from 1 to 10^2 and rarely more.

Let us now turn our attention to stage selectivity. It is necessary first of all to determine the stage bandwidth. To do this, we find the equivalent Q -factor of the stage as a whole; it is defined as

$$Q_{eq} = \frac{1}{G_{total}\rho} = \frac{1}{(G_c + G_i + G_g)\rho}$$

If conductances G_i and G_g are small as compared with the resonance conductance of the tuned circuit, then $Q_{eq} \approx Q_c$. Hence, the bandwidth will be

$$B = \frac{f_0}{Q_{eq}} = f_0 G_{total} \rho \quad (12-11)$$

The above discussion confirms the advantages of the pentode for amplification of radio signals.

If we take a stage having quality factor $Q_{eq}=80$, then at frequency $f_0=80$ kHz ($\lambda=3750$ m), the bandwidth $B=f_0/Q_{eq}=1000$ Hz. Such a bandwidth is completely unsuitable for radiotelephone reception. In fact, waves exceeding 2000 m are not used for radiotelephone transmission. If frequency $f_0=800$ kHz ($\lambda=375$ m), then with the same circuit Q -factor, we obtain the bandwidth $B=10,000$ Hz which is quite enough for broadcast reception. Finally, if we take signal frequency $f_0=8$ MHz ($\lambda=37.5$ m, i.e. in the short-wave range), the bandwidth B will be equal to 100,000 Hz which is clearly too wide for radiotelephony and impairs receiver selectivity.

This means that in some instances the bandwidth is deliberately extended at long waves as compared with its attainable value. At short waves the bandwidth obtained proves to be always wider than it is required from the viewpoint of radiotelephone reception selectivity. It is precisely these properties of tuned circuits and tuned amplifiers that make it necessary for the superheterodyne reception to be used, i.e., the reception followed by the conversion of signal frequency to the intermediate frequency which ensures the desired bandwidth when the Q -factor of the tuned circuits is normal. In radiotelephony the intermediate frequency is usually 465 kHz ($\lambda \approx 645$ m). Signal spectra are much wider in FM broadcasting or television reception and, therefore, the intermediate frequency in these cases is selected in other bands.

The form factor of a tuned stage as of any other single-circuit system equals 10 per cent. The form factor grows with the increasing number of tuned stages.

Directing our attention to the evaluation of amplification non-uniformity within the band used as the circuit is retuned with the aid of a variable capacitor, we shall use formula (12-10) and assume approximately that $G_{total} \approx \approx G_c = 1/Q_c \rho$. Thus, we may write

$$K_0 \approx S Q_c \omega_0 L \quad (12-12)$$

When retuning the circuit, only frequency ω_0 may be considered to be variable. Consequently, gain increases over one band *proportionally to the frequency*. When changing

over to the subband of shorter waves, L is switched to a lower-inductance coil, thus decreasing the average gain. This is illustrated in Fig. 12-14, where the resonance gain is shown as a function of the resonance frequency of a receiver for two subbands. Receiver gain is estimated, of course, at the worst point and, therefore, an appreciable non-uniformity within the subband and when changing over the subbands is harmful.

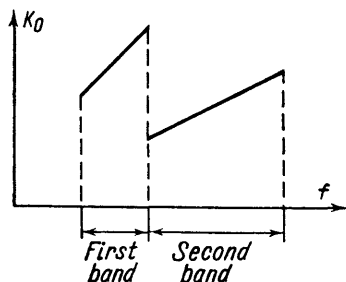


Fig. 12-14. Gain non-uniformity within two subbands

Gain non-uniformity may be compensated by providing the opposite non-uniformity for transmission gain in the input circuit (see Fig. 12-8). In the case of several stages, a uniform total gain ($K_{total} = K_{01} \cdot K_{02} \dots$) may be obtained

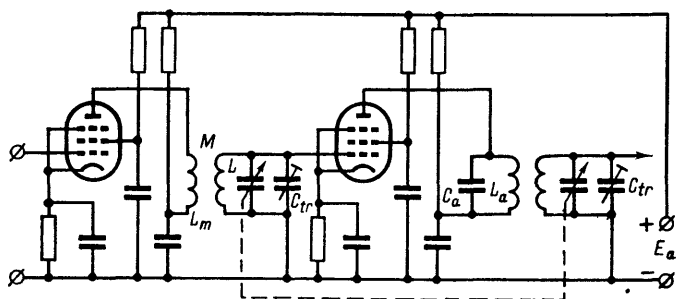


Fig. 12-15. Two-stage tuned amplifier employing the pentodes

by using different stage designs. Moreover, a multistage amplifier possesses a higher selectivity.

Figure 12-15 illustrates a schematic diagram of a two-stage tuned amplifier suitable for medium waves. As distinct from the diagram in Fig. 12-13, tuned circuits and anode circuits are inductively coupled in this amplifier. In this case, capacitors C_{block} and resistors R_g are not required. Tuning of the circuits in both stages is accomplished with the aid of ganged capacitors whose rotors are mounted on a common

axle (shown by the dashed line). Tuning may be adjusted with the aid of parallel-connected trimmer capacitors C_{tr} .

In the anode circuit of the 1st stage only coupling inductor L_m is connected, which has the coefficient of mutual inductance M with the tuned circuit inductor L . Referring to the relation $m=M/L$ as a transformation ratio, we may obtain the following formula for the gain of such a stage:

$$K_0 \approx SmQ_c\omega_0L \quad (12-13)$$

In practice, $m < 1$; therefore, the gain is reduced as compared with that determined by formula (12-12). Gain reduction may be required for the following reasons—first, to level up an average gain K_0 between subbands; second, to ensure stable operation of the amplifier; third, to reduce a shunting effect of valve output conductance on the tuned circuit.

The second stage differs from the first by the presence of capacitor C_a connected in parallel with anode inductor L_a . The function of this capacitor is to tune the anode circuit to the frequency $\omega_a < \omega_{0\min}$ and to level up gain within the subband similarly to the inductive-coupled input circuit (see Fig. 12-8).

Let us now consider *transistor tuned amplifiers*. HF transistors are characterized by their high cutoff frequency and low intrinsic feedback conductance. The common-emitter circuit is the basic circuit used in tuned stages.

Figure 12-16 illustrates two versions of transistor tuned stages. In the first one (Fig. 12-16a), the collector and output are autotransformer-coupled to the tuned LC circuit, with transformation ratios being m_1 and m_2 , respectively. Such connections are absolutely necessary, especially if we take into account that the output conductance of any given transistor and the input conductance of the succeeding one are extremely high. In case of direct (full) connection of the circuit, these conductances would severely shunt the tuned circuit so that it would lose its selectivity. The tuned circuit consists of inductor L (with taps m_1 and m_2), variable capacitor C and series-connected capacitors of comparatively high capacitance C_1 and C_2 , with capacitor C_1 being used to protect the inductor against possible short-circuit between the plates of capacitor C . Other parts of the circuit are well-

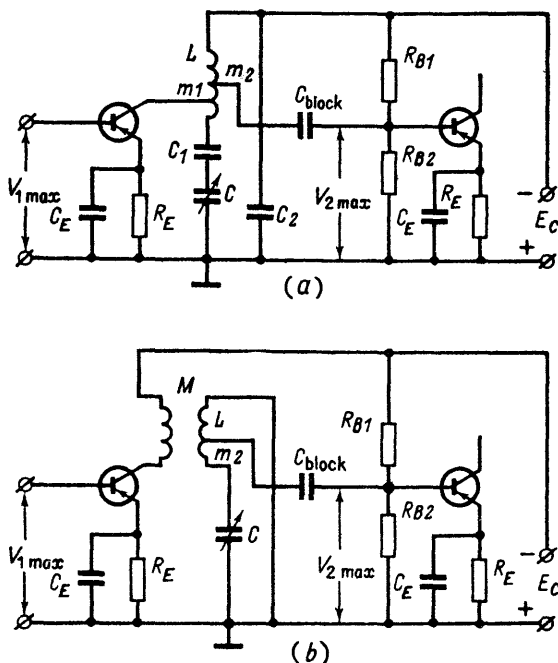


Fig. 12-16. Tuned amplifier stages employing transistors

(a) with double autotransformer connection of tuned circuit; (b) with inductive coupling between tuned circuit and collector circuit, and autotransformer connection of succeeding stage

known from Chapter IX (capacitance C_{block} , of course, is required of much lower value than in LF amplifiers). The second version of the stage (Fig. 12-16b) has a collector circuit inductive-coupled to the tuned circuit ($m_1 \approx M/L$) and an autotransformer output ($m_2 \approx L_2/L$, where L_2 is the inductance of turns between the base terminal and the stage chassis). In this version there is no need for capacitors C_1 and C_2 since the tuned circuit is not connected with a supply source.

By using the above method of reduction of circuit parameters to circuit terminals (see formula 12-9), we shall perform such a reduction and draw the resultant equivalent circuit (Fig. 12-17). In this equivalent circuit the current of signal source SV_1 reduced to tuned circuit terminals

should be multiplied by m_1 so that supply power remains unchanged. The output voltage reduced to the circuit terminals should be divided by m_2 taking into account voltage step-down transformation at the output. In the case of direct connection of the tuned circuit, corresponding conductances should be multiplied by m_1^2 and m_2^2 , respectively. Assume that the output conductance equals g_{22} and the input conductance of the succeeding stage is approximately equal

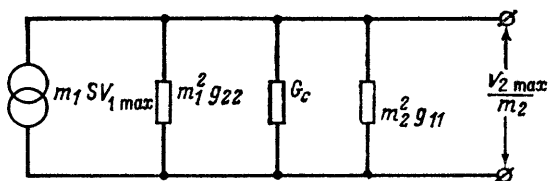


Fig. 12-17. Equivalent circuit of transistor stage at resonance

to g_{11} . At parallel resonance the inductive susceptance and the susceptances of the capacitive branches cancel each other. Assuming that the condition of resonance at the signal frequency is fulfilled, we omit the reactive branches in the schematic diagram shown in Fig. 12-17.

Let us examine stage gain and selectivity. In compliance with a general rule the gain for this equivalent circuit will evidently be

$$K_{0\,eq} = \frac{S}{G_{total}} = \frac{S}{m_1^2 g_{22} + G_c + m_2^2 g_{11}} = \frac{V_{2\max}}{m_2} : m_1 V_{1\max}$$

where G_c is the parallel resonance conductance of the tuned circuit itself ($G_c = 1/Q_c \rho$).

The actual voltage gain of the stage is

$$K_0 = \frac{V_{2\max}}{V_{1\max}} = K_{0\,eq} m_1 m_2 = \frac{S}{G_{total}} m_1 m_2 \quad (12-14)$$

The equivalent quality factor of the stage (i.e., the tuned circuit shunted by reduced conductances) may be found from formula (12-8) in the form

$$Q_{eq} = \frac{1}{G_{total} \rho} = \frac{1}{\rho (G_c + m_1^2 g_{22} + m_2^2 g_{11})} \quad (12-15)$$

and the stage bandwidth will prove to be equal to

$$B = \frac{f_0}{Q_{eq}}$$

Gain non-uniformity within the band is similar to that of a valve stage employing direct (full) connection of a tuned circuit.

As we know a transistor stage may be more fully estimated by the value of its power gain. In general, the power gain may be expressed as

$$K_P = \frac{P_{out}}{P_{in}} = \frac{V_{2\max}^2 g_{out}}{V_{1\max}^2 g_{in}}$$

where g_{out} is the load conductance of the given stage and g_{in} , its input conductance. If we assume that $g_{out} = g_{11}$ and $g_{in} = g_{11}$ (i.e., that the given stage is identical with the succeeding one),

$$K_P = K_0^2$$

Finally, if we neglect losses in the tuned circuit itself, i.e. if we assume that $G_c = 0$ and match the given stage to the succeeding one by selecting transformation ratios from the condition that source and load conductances are equal,

$$m_1^2 g_{22} = m_2^2 g_{11} \quad (12-16)$$

i.e. we select

$$\frac{m_1}{m_2} = \sqrt{\frac{g_{11}}{g_{22}}}$$

then the voltage gain from formula (12-14) takes the form

$$K_{0\text{ avail}} = \frac{m_1 m_2 S}{m_1^2 g_{22} + m_2^2 g_{11}} = \frac{m_1 m_2 S}{2 m_1^2 g_{22}} = \frac{S}{2} \frac{m_2}{m_1 g_{22}} = \frac{S}{2 \sqrt{g_{22} g_{11}}} \quad (12-17)$$

Available power gain (under matching conditions) is

$$K_{P\text{ avail}} = K_{0\text{ avail}}^2 = \frac{S^2}{4 g_{22} g_{11}} \quad (12-18)$$

It is this, so to say, idealized parameter that may be used to estimate the transistor itself. For instance, if it is

known that $g_{22}=1\text{ mS}$, $g_{11}=4\text{ mS}$ and $S=30\text{ mS}$, then $K_{P_{avail}}=30^2/4 \times 1 \times 4 \approx 56$ (17.5 dB).

It will be recalled once again that the high-frequency parameters of a transistor differ from the low-frequency ones. In addition, good handbooks furnish the relationship between these parameters and frequency. If a stage, however, is designed for amplification at frequencies which are much lower than the cutoff frequency of the transistor used, we may use, as said earlier in Section 9.2, low-frequency values of these parameters.

12-4. INTERMEDIATE-FREQUENCY AMPLIFIERS

IF amplifiers in superheterodyne receivers also contain tuned circuits in their stages. But as distinct from the retunable stages described above, an IF stage features fixed circuit tuning and this tuning corresponds to the intermediate frequency f_{IF} specified for a given receiver. As we have already mentioned, the intermediate frequency for broadcast receivers is generally chosen to be 465 kHz ($\lambda \approx 645\text{ m}$). In the case of fixed tuning the possibility exists of connecting into each stage not one but two or even more tuned circuits forming a bandpass filter. Such IF amplifiers are known as *bandpass amplifiers*. They are capable of ensuring better selectivity than single-circuit amplifiers.

Figure 12-18a illustrates a valve IF amplifier stage with a two-circuit filter. The first LC circuit of this filter is connected directly to the anode circuit, and the second one is inductively coupled with the first circuit and connected to the succeeding stage input. Inductors L are small in size and are provided with magnetic cores which make it possible to adjust circuit tuning. Both circuits are tuned to the same intermediate frequency.

Bandpass stage gain depends not only upon valve and circuit parameters but also upon the degree of coupling between the tuned circuits. With critical coupling, i.e., when the degree of coupling is such that the resistance inserted from the second circuit to the first one equals to the intrinsic resistance of the latter, the gain at the resonance frequency will be twice as less as that of a single-circuit stage, since the signal power is distributed equally between the two tuned

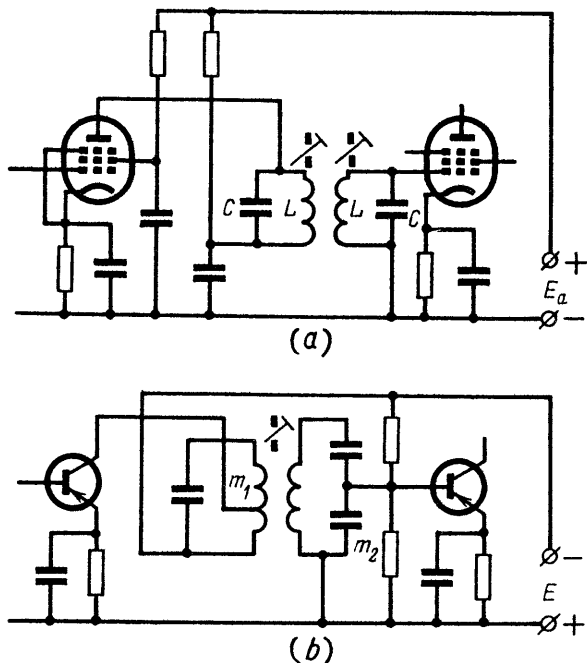


Fig. 12-18. IF bandpass amplifiers
 (a) employing the pentodes; (b) employing the transistors

circuits. Proceeding from formula (12-10), we may write

$$K_{0 \text{ crit}} \approx \frac{S}{2G_{\text{total}}} = \frac{SQ_{eq}\omega_0 L}{2} \quad (12-19)$$

Here, Q_{eq} is the equivalent quality factor of each tuned circuit with allowance made for the valve shunting effect but without accounting for the mutually inserted resistances. Equivalent Q -factors of both tuned circuits may be considered approximately equal in magnitude.

For instance, if the equivalent resonance conductance of each tuned circuit $G_{\text{total}} = 0.1$ mS and valve transconductance $S = 4$ mS, then the voltage gain will be $K_{0 \text{ crit}} = 20$. Note that a critical value of the coupling coefficient is determined from the relation $k_{\text{crit}} = 1/Q_{eq}$; $k_{\text{crit}} = 2\%$ when $Q_{eq} = 50$.

To analyze selectivity of a bandpass amplifier, let us examine Fig. 3-6 from the chapter on coupled circuits. As noted in this chapter, each resonance characteristic curve of two coupled circuits has a more flat top and more steep edges than a single circuit. In particular, when coupling is more than critical, resonance characteristic curves have a double-humped form, and the stage bandwidth is approximately equal to the difference between the resonance frequencies of the two circuits given by formulas (3-15) and (3-16). Form factors of such stages depend upon the degree of coupling but they, of course, are much greater than for a single circuit (i.e., more than 10%). An example of a double-humped characteristic curve is shown in Fig. 12-19. The curve represents the relationship between the relative gain and the detuning increment.

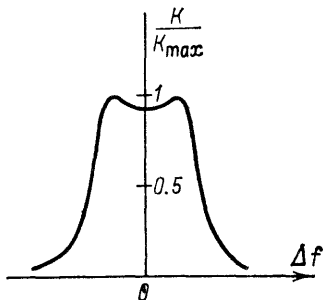


Fig. 12-19. Frequency response of bandpass amplifier stage

The double-humped resonant characteristic, however, is not widely used since it complicates a tuning procedure. In addition, in the case of tight coupling and when many stages are used, the valley at the center frequency may become deeper, thus impairing the amplification of the carrier. It is common practice to select critical coupling (or a value close to it). With the critical coupling between two circuits the stage bandwidth is

$$B_{crit} = \frac{\sqrt{2}f_0}{Q_{eq}} \quad (12-20)$$

i.e., is 1.41 times greater than for a single circuit having the same quality factor. In turn, the form factor of such a stage characteristic curve will be

$$\frac{B_{crit}}{2\Delta f_{0.1\ crit}} \approx 0.32 \quad (\text{or } 32 \text{ per cent})$$

This is the major advantage of a bandpass stage for the sake of which we sacrifice the value of voltage gain. Never-

heless, in modern valves having a high transconductance the desired gain is obtained with a small number of IF stages (one or two, but seldom three stages in broadcast receivers).

Figure 12-18*b* illustrates a schematic diagram of an IF amplifier transistor stage employing a two-circuit filter. Characteristic of these circuits is the fact that to retain their high quality factor it is necessary to reduce the shunting effect of the stage output and input conductances by employing a partial connection of the circuits (tapping factors m_1 and m_2).

By analogy with formulas (12-14) and (12-19), stage gain at the critical coupling between the circuits will be defined as

$$K_{0crit} = \frac{S}{2G_{total}} m_1 m_2 = \frac{SQ_{eq}\omega_0 L m_1 m_2}{2} \quad (12-21)$$

Values of m_1 and m_2 are selected by considering the admissible deterioration of Q -factors compared to those of the tuned circuits themselves as shown, for instance, by formula (12-9).

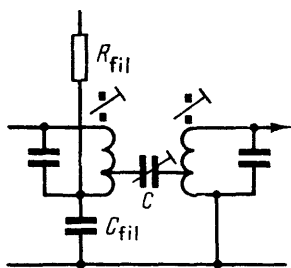


Fig. 12-20. Bandpass filter with two capacitively-coupled circuits

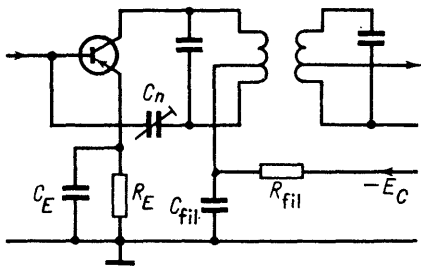


Fig. 12-21. Bandpass amplifier stage with compensation of capacitor C_{12}

At the critical coupling the bandwidth of the transistor stage, just like that of the valve stage, will be determined by formula (12-20), and its selectivity will also be identical to the latter. One of the parameters used in selecting the type of transistor for a bandpass stage is the idealized power gain K_{Pavail} defined by formula (12-18). This performance characteristic for a two-circuit stage proves to be the same as for a single-circuit stage since here we neglect losses in the tuned circuits.

Apart from the above stages employing bandpass filters in the form of inductively coupled tuned circuits there are stages with bandpass filters whose tuned circuits are coupled through an "external" capacitor (Fig. 12-20).

In this case, each circuit may be tuned to the desired frequency by the magnetic core while the coupling may be selected by capacitor C (with a capacitance usually equal to several picofarads).

In transistor amplifiers (especially in those employing transistors with a cutoff frequency not exceeding much the value of intermediate frequency), there exists the intrinsic feedback capable of disturbing the desired performance of a stage. Therefore, compensation of this feedback is sometimes accomplished in such amplifiers. For this purpose, a voltage, equal in magnitude but opposite in phase to that fed back to the input through the intrinsic capacitance C_{12} , is applied from the output to the input through a compensating capacitor C_n (Fig. 12-21).

The opposite phase of the compensating voltage is obtained by connecting the compensating circuit to the opposite end of the 1st circuit's inductor having an intermediate lead (tap) to the collector supply source. In HF transistors reverse transfer conductance g_{12} is, as a rule, low and does not require compensation.

12-5. RADIO INTERFERENCE

Any undesired electric (or magnetic) energy that tends to interfere with the reception of desired signals by producing sound (noise) in the loudspeaker, errors in telegraph records or stripes and flashes on the TV screen are referred to as RF interference.

Among the sources of RF interference may be undesired transmitters, electric processes of natural origin, electrical equipment, fluctuations of electrons in receiver valves and circuits. By the waveform of the voltages generated in the radio receiver interference may be classified under three types: periodic, pulsed and flat.

The most serious complications for radio reception, especially at short waves, are caused by the interference produced by undesired transmitters since the number and power of

radio stations are continuously increasing all over the world and short waves are capable of propagating over great distances. Interference of this type has usually a periodic waveform, i.e., a definite frequency or more precisely a definite frequency spectrum. Therefore, the measures for increasing receiver selectivity described above are provided precisely to eliminate interference from undesired transmitters. In

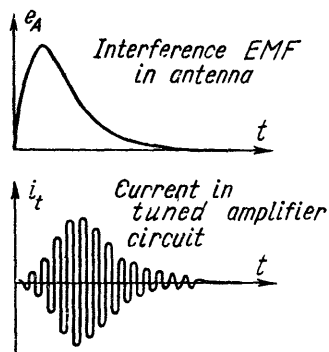


Fig. 12-22. Pulse interference

general, these measures are reduced to obtaining high rectangularity of a receiver's resonant characteristic curve (see Fig. 12-3) limited by the bandwidth corresponding to the width of the signal spectrum.

Directional antennas may serve as an additional means of increasing receiver selectivity. It is common practice in special radio communication to use directional reception (just as directional transmission)

between two stationary stations. It is also possible to employ a magnetic antenna whose directivity pattern maximum is oriented on the station being received or whose directivity pattern minimum (zero) is directed towards the station which produces the most hazardous interference.

Among the natural kinds of interference the first place is taken by *atmospheric discharges* which manifest themselves as short cracks or more prolonged rustles in the loudspeaker. Such interference is caused by lightning in the vicinity of a receiver or produced by the field propagating from tropical areas. The level of atmospheric interference is reduced on average with increasing geographical latitude. In summer atmospheric interference is much severe than in winter.

The EMFs induced by atmospheric discharges in a receiving antenna present basically short pulses or bursts capable of producing damped oscillations at the natural frequency of the circuit (Fig. 12-22). Since such a pulse does not have its own carrier frequency, one cannot speak about tuning or detuning of a receiver relative to atmospheric inter-

ference. The interference level, however, depends essentially upon the frequency band of the receiver. Atmospheric interference produces the maximum effect at long waves, whereas at ultrashort waves it does not practically disturb radio reception.

This may be explained by the following physical property of pulse interference. If an aperiodic pulse is expanded into components, we obtain an infinite spectrum consisting of oscillations of different frequencies, from the lowest to the highest. In this spectrum, however, those components whose cycles are close to the pulse duration will have relatively greater energy. By considering that the pulse of atmospheric interference lasts for dozens of microseconds one can easily understand why such interference is most hazardous at long waves, which is well known to radio amateurs.

Among the natural kinds of interference are also voltage pulses induced in the receiving antenna by blows of snow flakes, particles of sand (electrification by friction) and antenna electrification by a static atmospheric field.

Interference produced by electrical equipment located close to a receiving antenna is known as *man-made interference*. This interference appears as a result of abrupt changes in electric and magnetic fields induced by such equipment when their fields reach the receiving antenna. Fields are changed at current commutation in electrical equipment, especially if the commutation is accompanied by arcing. Consequently, among the sources of man-made interference are, for instance, commutator motors, electric relays and bells, breakers, telegraph and medical equipment, etc.

EMFs induced in the receiving antenna by these interference sources are also essentially of a pulse nature. Consequently, one cannot tune away from this interference but the maximum of interference energy may lie within a particular wave range which is determined by the pulse duration. For instance, ignition interference is noticeable in the television broadcast band.

Among essential measures taken for reduction of man-made interference are the shielding of these interference sources to prevent the action of their fields on the receiving antenna and the RF filtering (by-passing) of power supply circuits to prevent interference currents from entering the receiver thro-

ugh power-supply connecting wires (Fig. 12-23). Interference level in cities should be controlled by legislation. In rural areas the level of man-made interference is much lower than in a city.

At ultrahigh frequencies (especially at decimetric and centimetric waves) atmospheric and man-made interference are not critical from the viewpoint of determining the required signal field strength which ensures superiority of the signal over interference. The minimum signal field strength

required for reception is determined by the *noise* arising as a result of random motion of electrons in the vacuum valves, semiconductor devices and circuits of the receiver itself.

Suppose that we have decided to increase receiver sensitivity so that to make the receiver more sensitive to the signals with

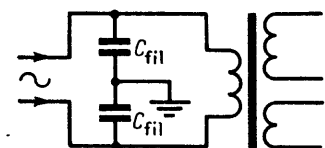


Fig. 12-23. Reduction of man-made interference in receiver power transformer

smaller amplitudes. With this aim in view, we increase the number of amplifier stages. If we could manage to eliminate the effect of feedback and the action of external interference, we should come sooner or later to the level of sensitivity below which the signal would be suppressed by internal noise of the receiver.

Ordinary resistors in amplifier input circuits are one of the sources of noise voltage. We are accustomed to think that a resistor is a passive element which produces no oscillations. In fact, however, electrons ("free" electrons, i.e., conduction electrons) move at random inside the resistor material. The higher the resistor temperature, the more intensive are these motions. Random electron current produces AC voltage with arbitrary amplitudes and frequencies across resistor R .

The average magnitude of this random voltage is zero. These voltage variations (called *fluctuations*) act on the amplifier input. If fluctuation peaks are comparable with input signal amplitudes, the output signal may prove to be indistinguishable at the background of the internal noise being perceived by the ear as hum.

With a high gain from the receiver input to a loudspeaker internal noise may be heard. This noise is of a hissing type,

i.e., it is a sound with no definite audio frequency, thus differing from tone.

From the spectral viewpoint, noise is a random sum of an infinite number of elementary components with arbitrary frequency relationships and accidental phase relationships. The noise spectrum covers all RF bands. Noise power is distributed uniformly over the frequency scale unlike the power of pulse interference. However, internal noise generally does not play a dominant role on the background of external noise in long-wave and medium-wave broadcast receivers having a comparatively narrow bandwidth. Thus, measures have to be taken for reducing noise only at ultrahigh frequencies (microwaves).

Other sources of fluctuations limiting the sensitivity of receiving equipment are fluctuations of currents inside the valve and semiconductor devices, especially in the input stage. Fluctuations of anode current in a valve are primarily caused by a random electron escape from the cathode. On the average, anode current, in the absence of a signal, is constant during finite time intervals. However, when comparing the quantity of electrons being emitted by the cathode for identical extremely short intervals of time they prove to be somewhat different. This means that anode current fluctuates near its mean value, and these fluctuations are capable of causing noise voltage at the input of the next stage. Similarly, we cannot consider collector current in a transistor to be constant.

The higher the temperature of noisy resistors, the greater the noise level. At room temperatures it is difficult to control this factor. The noise level (power) increases *proportionally to the bandwidth* since the spectrum of fluctuation frequencies passing through this bandwidth extends with the extending bandwidth. Consequently, noise is more hazardous for television than for radio broadcast reception. In valves the noise level depends upon the *number of electrodes*, i.e., the greater the number of grids, the greater the number of causes for electron flow fluctuations in a valve. From this viewpoint, it is more advantageous to employ a triode than a pentode in the receiver input stage. The high interelectrode capacitance of a triode, however, causes such a strong feedback that natural oscillations occur in the stage, thus making it impossible to amplify the signal.

In order that triodes could be employed in tuned amplifiers (especially at ultrashort waves), they are connected into the so-called *cascode amplifier** (Fig. 12-24). Two valves of this circuit are connected in series with respect to an anode supply source. The first stage contains valve $V1$ employing a grounded-cathode circuit. In the second stage the grid of valve $V2$ is connected (for alternating current) to a ground (to the amplifier chassis) through capacitor C . Consequently, the second valve employs a grounded-grid circuit. It is this combination of grounded-cathode and grounded-grid stages

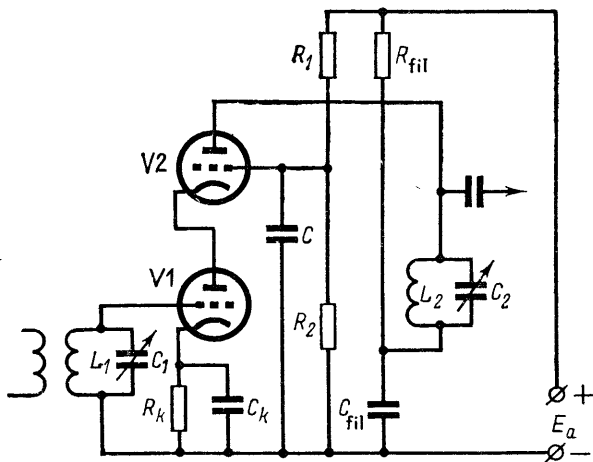


Fig. 12-24. HF cascode amplifier

that gives the indication of a cascode amplifier. The bias voltage applied to the 1st valve's grid is produced by anode current flowing through resistor R_k , and the bias voltage across the 2nd valve's grid with respect to the cathode is drawn from anode supply source E_a through voltage divider R_1R_2 .

Thus, the input (cathode-grid terminals) of the 2nd valve is connected between the anode-cathode terminals of the 1st valve and serves as a load. The input impedance of a grounded-grid stage (just as a common-base transistor stage) is

* A combination of grounded-cathode and grounded-grid stages is referred to as a "cascode amplifier".

very low. It is equal to $R_{in2} \approx 1/S$, where S is the valve transconductance. Consequently, the input impedance amounts to 250Ω at a transconductance, for instance, of 4 mS. Physically, this is explained by the fact that anode alternating current flows through the valve input circuit, thus reducing the input impedance.

With such a low load resistance, the 1st valve will retain normal operation irrespective of the high transfer capacitance of the triode. In this case, the gain of the 1st stage will not exceed unity ($SR_{eq} = SR_{in2} = S/S$). However, the input impedance of the 1st valve shunting the L_1C_1 circuit proves to be high and, therefore, does not impair the circuit quality factor.

The 2nd valve has a normal gain since it is loaded with the output circuit, L_2C_2 . Feedback that may appear in this valve through an anode-cathode capacitance is eliminated by the grounded grid functioning as a screen located between the anode and cathode. By this means, the cascode amplifier assures reduced noise power, high input impedance and gain which corresponds to the amplification coefficient of one valve. This circuit may employ both a double triode and transistors.

Noise in the receiver begins with the antenna. To the noise produced by the antenna resistance is added the noise induced in the antenna by earth radiation and cosmic radiation. Then the total noise power is supplemented by the noise produced in the receiver circuits and electron devices.

The concept of a noise factor is introduced to estimate how additional sources of noise impair the performance of an ultrahigh-frequency receiver. The receiver *noise factor* is the number indicating how many times the output noise power is greater than the noise power produced by the antenna resistance only.

$$N = \frac{P_{n \text{ out}}}{K_P P_{nA}}$$

The noise factor is always greater than unity. When amplifier noise performance is improved, this factor tends to unity.

There are some methods of decreasing the noise power with the aid of cryogenic engineering when the temperature

of input circuits is reduced to the value approaching the absolute zero (in practice, below -200°C). In this case, instead of conventional valves or transistors use is made of masers and parametric amplifiers as active devices. These amplifiers find very important applications in a number of electronics fields, such as space communication, radioastronomy, radiolocation, etc. In television receivers cooled amplifiers are not yet used.

12-6. FREQUENCY CONVERTERS

Let us return to the superheterodyne block-diagram given in Fig. 12-6. We have already mentioned that the purpose of frequency conversion is to select the intermediate frequency at which the bandwidth required for the given frequency band

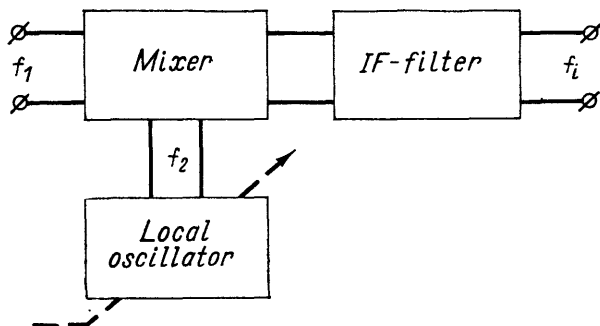


Fig. 12-25. Block diagram of frequency converter

could be obtained with normal values of the circuit quality factor.

The superheterodyne principle allows one to provide high sensitivity owing to the amplification in three regions of frequencies and the selection of high-quality circuits. High selectivity can be obtained since fixed tuning of IF amplifier stages allows bandpass amplifiers to be used. The basic amplification of a signal is performed at a fixed intermediate frequency and, therefore, sensitivity remains essentially unchanged when tuning the receiver or changing over its subbands. Finally, the high gain IF amplifier delivers to the

detector input the signal power required for successful detection and control of automatic circuits (automatic gain control, visual-tuning indicator, automatic tuning, etc.).

Thus, the most important process occurring in a superheterodyne is the *conversion of signal frequency* to intermediate frequency. This is accomplished in a frequency converter whose block diagram is given in Fig. 12-25. The block diagram

consists of a local oscillator (LO), i.e., a low-power self-excited oscillator which generates continuous waves at natural frequency f_2 , a mixer to which a signal of frequency f_1 and LO signal are applied simultaneously, and, finally, an IF filter. The principle of frequency conversion known from

Chapter XI is as follows. The mixer should be a non-linear device which produces oscillations of a sum frequency $f_2 + f_1$ and a difference frequency $f_2 - f_1$ under the action of the oscillations with frequencies f_1 and f_2 . Consequently, the term "mixer" is not apt as the process occurring in it is not the usual mixing of applied oscillations but the generation of oscillations with new frequencies. Such a mission may be accomplished by a non-linear device (a diode, amplifier valve or transistor).

The intermediate frequency f_i in a superheterodyne is determined (or, so to say, fixed) by the tuning of a filter connected to the mixer output. Intermediate frequency is obtained as the *difference* between the frequencies of the local oscillator and the incoming signal ($f_i = f_2 - f_1$). The oscillations of the sum frequency are not used. The tuning of the local oscillator would be double-valued if it were tuned independently of the preselector (the *preselector* is a channel operating at the frequency of the incoming signal, i.e., the input circuit and the tuned amplifier). In fact, the same IF may be obtained at the "upper" ($f_i = f_2 - f_1$) and "lower" ($f_i = f_1 - f_2$) tuning of the local oscillator. These possibilities are illustrated in Fig. 12-26 where the values of frequency f_2 are plotted on the abscissa as the local-oscillator frequency is

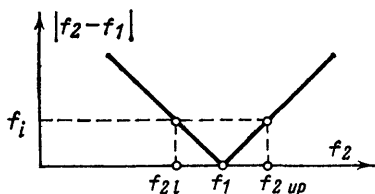


Fig. 12-26. Lower and upper tuning of local oscillator

varied and frequency f_1 of the received signal is marked, whereas corresponding absolute values of the difference frequency, including the fixed intermediate frequency f_i , are laid off on the ordinate. The value of intermediate frequency f_i is obtained when the local oscillator frequency is varied to the left and right from resonance frequency. In other words, "upper" ($f_{2up} = f_1 + f_i$) and "lower" ($f_{2l} = f_1 - f_i$) tuning of the local oscillator are equally useful.

Superheterodyne tuning should be accomplished with the aid of one knob (ganged tuning). For this purpose, the capacitor of the LO circuit is mounted on a common axle with preselector capacitors. When tuning, LO frequency should be changed simultaneously with preselector circuit frequency so that at each point of the scale the frequency difference will be constant and equal to f_i . If this is the case, one should select either the upper or lower tuning of the local oscillator and carry out tracking accordingly. The upper tuning which we shall discuss here is very convenient in broadcast bands. The lower tuning is the one most frequently encountered in ultrahigh-frequency (microwave) receivers where it is easier to select an active device and stabilize LO frequency at a reduced frequency.

This desired relationship between the RF-signal circuit (f_1) and LO circuit (f_2) tunings and the angle of turn of identical straight-line-frequency variable capacitors may be presented graphically as shown in Fig. 12-27a. In this figure both tunings are represented by parallel lines spaced vertically by f_i . For the signal frequency when identical capacitors C are used, a reduced inductance ($L_{LO} < L$) in the LO circuit has to be taken, hence

$$\frac{1}{2\pi \sqrt{L_{LO}C}} > \frac{1}{2\pi \sqrt{LC}}$$

Even if this condition is satisfied, we shall not obtain the desired result since LO frequency will follow the dashed (steeper) line when tuning over the tuning range, and the proper frequency difference will be obtained only at one point (in our example, at point a , Fig. 12-27).

In order to reduce the tracking error, i.e., the discrepancy between the solid and dashed lines corresponding to frequency f_2 , the LO circuit has a series capacitor C_2 and a parallel

capacitor C_1 added (Fig. 12-27b). At the beginning of the scale (at lower frequencies) the capacitor C_2 reduces the large capacitance of capacitor C , thus increasing the frequency so that the dashed line f_2 coincides with the solid one. At the end of the scale (at higher frequencies) capacitor C_1 has a pronounced effect on circuit capacitance by increasing it, thus decreasing the frequency so that the dashed line f_2 coincides with the solid one. By this means, accurate tracking is achieved.

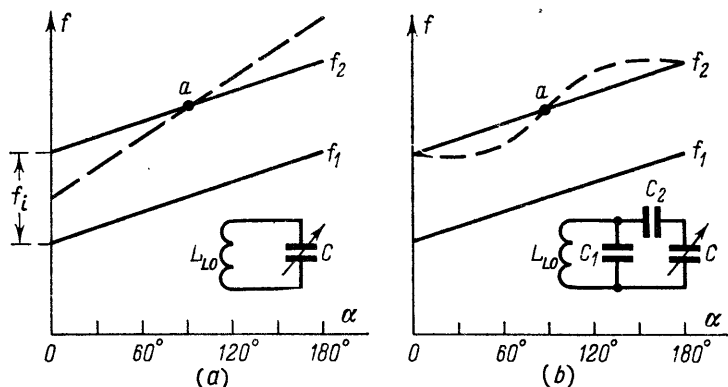


Fig. 12-27. Principle of local-oscillator tracked tuning

(a) tuning error in case of simple circuit; (b) tuning errors in case of complex circuits

ved at three points of the tuning range (subband) and the tracking error is reduced at other point. Such a complicated circuit of the local oscillator we encounter in many frequency converters. Incidentally, tracking at one or two points only is proved sufficient in narrow shortwave bands with extended scales.

Frequency converters for long, medium and shortwave receivers employ primarily special converter valves (see Chapter VIII). Among the advantages of such valves are not only economies in the number of receiver valves but advantageous conditions are provided for interaction between the local oscillator and the mixer inside the electron valve. An example of a frequency converter employing a heptode is shown in Fig. 12-28. In this diagram an LO circuit is coupled to the 1st (from the cathode) valve grid and has a three-point type cat-

hode feedback. The LO circuit consists of inductor L_{LO} and capacitors C , C_2 and C_1 . One may say that all the electrodes at positive potential (the 2nd grid, 4th grid and the anode proper) serve as an anode of the LO valve. The signal from the local oscillator participates in the control of a common electron flow in the valve, thus maintaining the voltage of frequency f_2 between the 1st grid and the cathode.

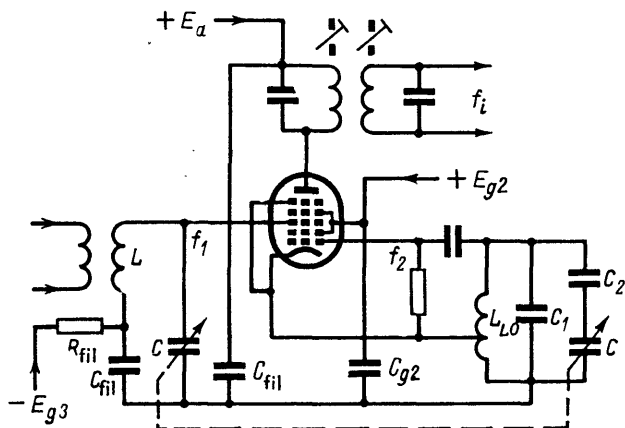


Fig. 12-28. Two-grid frequency converter employing a heptode

The signal voltage fed from the LC circuit of the receiver's tuned amplifier (or input circuit) to the 3rd grid of the heptode also participates in the control of electron flow. The 4th and 2nd grids serve as screen grids, whereas the 5th grid is a suppressor one. Not only currents at frequencies f_1 and f_2 but current at a difference frequency $f_i = f_2 - f_1$ are produced in the anode circuit owing to variations in transconductance. The difference-frequency current feeds a two-circuit bandpass filter tuned to the required frequency f_i . Intermediate-frequency voltage is taken from the filter output to the input of a succeeding amplifier stage. Sometimes, this kind of conversion is referred to as "two-grid" conversion in order to underline the action of signal- and LO voltages on different grids of the valve. Bias is applied to the signal (3rd) grid from supply source E_{g3} through filtering element ("isolator") $C_{fil}R_{fil}$.

Let us explain the operating principles of a two-grid frequency converter and give a quantitative estimate of conversion results. As a matter of fact, single-grid conversion has been explained in Chapter XI by non-linearity, i.e., by variations in the slope (transconductance) of characteristic under the combined action of signal- and LO voltages. In our case these voltages act on different electrodes and, therefore,

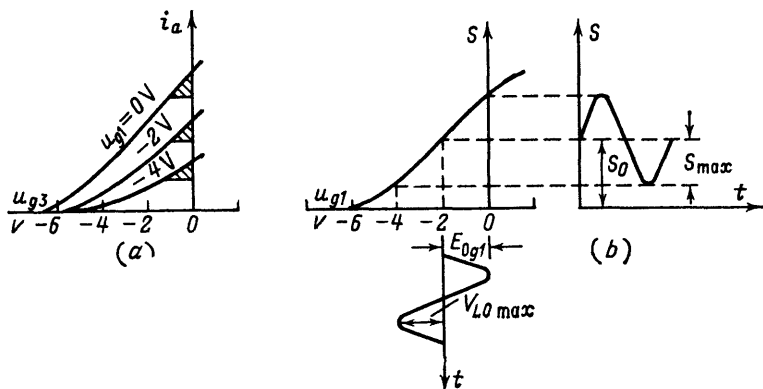


Fig. 12-29. Illustration of two-grid frequency conversion

(a) anode current as a function of signal-grid voltage for different voltages applied to local-oscillator grid; (b) transconductance with respect to signal grid as a function of local-oscillator voltage

it is not quite evident that transconductance is dependent upon LO voltage.

Let us consider the relationship between anode current i_a and voltage u_{g3} across the signal (3rd) grid to be the major characteristic of a heptode. This characteristic is similar to the conventional anode-grid characteristic of a pentode. However, if such characteristics are taken at different potentials u_{g1} of a local oscillator (1st) grid, we shall obtain a "fan-shaped" family of characteristic curves (Fig. 12-29a): *the higher the potential of the LO grid, the steeper the anode-signal grid characteristic curve*. Physically, the increase in the slope of the anode-current characteristic curve is explained by the presence of two space charges in a heptode — near the 1st and 3rd grids. When electron flow passing through the 1st grid grows, the upper space charge controlled by signal grid

voltage also increases, and the anode current increases more sharply when this voltage becomes higher. The shaded triangles have vertical sides whose length is proportional to the transconductance. By using these sides, we may plot the *relationship between the transconductance with respect to the signal grid and the voltage across the LO grid* (Fig. 12-29b). Assume, for instance, that this relationship corresponds to the signal grid voltage $u_{g3}=0$. Let the local oscillator produce AC voltage $V_{LO \max}$ at frequency f_2 (with an amplitude of 2 V) and the grid bias (of about 1 V) across the 1st grid; this is illustrated as a function of time with the time axis being directed downwards. It is evident that the transconductance with respect to the signal grid will vary periodically at the frequency f_2 and amplitude $S_{\max}$, which is also shown in the figure (to the right) as a function of time.

The instantaneous value (in time) of transconductance will consist of DC- and AC components, namely,

$$S = S_0 + S_{\max} \cos 2\pi f_2 t$$

If signal voltage $V_s \max \cos 2\pi f_1 t$ is simultaneously applied to the 3rd grid, an alternating current equal, as usual, to the product of the transconductance and AC grid voltage ($i_a = t V_s \max \cos 2\pi f_1 t$) will arise in the anode circuit. However, the transconductance itself varies with time and, therefore, the resultant current will be

$$i_a = (S_0 + S_{\max} \cos 2\pi f_2 t) V_s \max \cos 2\pi f_1 t$$

Opening the brackets we obtain

$$i_a = S_0 V_s \max \cos 2\pi f_1 t + S_{\max} V_s \max \cos 2\pi f_2 t \cos 2\pi f_1 t$$

The second term in the above expression is a cosine product with different arguments, which may be replaced by the following sum:

$$\cos \alpha \cos \beta = \frac{1}{2} \cos(\alpha + \beta) + \frac{1}{2} \cos(\alpha - \beta)$$

By substituting the values from our formula for the arguments, i.e.,

$$\alpha = 2\pi f_2 t; \quad \beta = 2\pi f_1 t$$

we obtain the final value for the anode current components being produced by the combined action of signal- and LO

voltages,

$$i_a = S_0 V_{s \max} \cos 2\pi f_1 t + \frac{1}{2} S_{\max} V_{s \max} \cos 2\pi (f_2 + f_1) t + \\ + \frac{1}{2} S_{\max} V_{s \max} \cos 2\pi (f_2 - f_1) t$$

This result confirms mathematically the physical processes discussed above. Note that along with the signal-frequency component (f_1) the anode current also contains the components with the sum frequency ($f_2 + f_1$), which we have mentioned briefly before, and the difference frequency ($f_2 - f_1$). It is precisely the *oscillation of the difference frequency which is the useful result of the conversion*.

The amplitude of difference-frequency current is

$$I_{i \max} = \frac{1}{2} S_{\max} V_{s \max}$$

where $S_{\max}$, as we remember, is the amplitude of transconductance variations under the action of LO voltage. $S_{\max}$ depends both upon the shape of the operating portion of the valve characteristic and the amplitude of LO voltage.

IF voltage across the terminals of the tuned circuit (in the case of a single-circuit load) is equal to the product of the current (at this frequency) and the resonance impedance

$$V_{i \max} = I_{i \max} Z_{c \text{ res}} = \frac{1}{2} S_{\max} V_{s \max} Z_{c \text{ res}}$$

The ratio of this "output" voltage to the signal voltage (the ratio known as the conversion efficiency) is

$$K_C = \frac{V_{i \max}}{V_{s \max}} = \frac{1}{2} S_{\max} Z_{c \text{ res}} = S_C Z_{c \text{ res}} \quad (12-22)$$

The value of $S_C = 1/2 S_{\max}$ is referred to as the *conversion transconductance*. This is a parameter which makes it possible to estimate a converter valve and its operating conditions. Conversion transconductance is also measured in millisiemens.

The value of this parameter should be as great as possible. For this purpose, the transconductance $S_{\max}$ should also be great but, as seen from Fig. 12-29, it increases with the increase in LO voltage $V_{LO \max}$. This means that the LO vol-

tage applied to the mixer input must also be sufficiently high (about 2 V in Fig. 12-29).

When LO voltage is increased, however, up to the values at which the anode current will be clipped, i.e., discontinued for some portion of a cycle, the increase in conversion transconductance will stop.

Approximately it is considered that the value of $S_c = S/4$ (where S is the transconductance corresponding to the linear portion of a pentode characteristic curve) is the maximum value of conversion transconductance.

Let us consider a numerical example. Assume that the pentode-mixer characteristic transconductance $S = 8$ mS, the IF tuned circuit coupled to the anode circuit has a resonance impedance $Z_{c\ res} = 5000\Omega$ and the signal voltage across

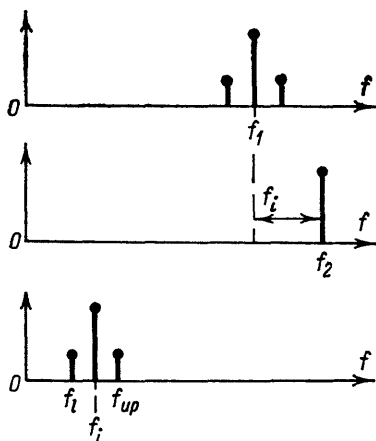


Fig. 12-30. Transfer of signal spectrum to intermediate frequency

the control grid $V_{s\ max} = 1$ mV.

If we have selected the best operating conditions for heterodyning, the IF voltage across the tuned circuit terminals will be

$$V_{i\ max} = V_{s\ max} K_C = V_{s\ max} S_c Z_{c\ res} = \frac{1}{4} V_{s\ max} S Z_{c\ res} = \\ = \frac{1}{4} \cdot \frac{8}{1000} 5000 = 10 \text{ mV}$$

One can see from this example that in comparison with frequency conversion with the use of a diode the conversion in an amplifier valve is more advantageous since along with signal-frequency conversion we have obtained a 10-time increase of signal voltage (i.e., $K_C = 10$).

Thus, a frequency converter employing an amplifier valve may be designed as an amplifier stage with a reduced

transconductance by taking the signal and IF oscillation amplitudes as input and output voltages, respectively.

There are several types of converter valves and converter circuits. The physical processes occurring in them and the design principles are similar to those laid down above.

Essential is the following property of frequency conversion — if a signal with a carrier frequency f_1 is modulated, its spectrum will be transferred, as a result of the frequency conversion, to a new carrier frequency f_i . This is easily proved by means of Fig. 12-30. The upper plot in this figure represents the modulated signal spectrum consisting of one carrier and two sidebands. In the middle plot the spectral line of LO oscillation is shown. Its frequency exceeds the signal frequency by f_i . The lower plot is the spectrum of difference frequencies where the central position is occupied by f_i ; the upper side frequency f_{up} is obtained as a result of subtracting the lower side frequency from the LO frequency, and the lower side frequency f_l is derived by subtracting the upper side frequency. In other words, the *law of signal modulation is retained* after frequency conversion.

In those possible frequency converters where the LO frequency f_2 is lower than the signal frequency f_1 , the upper and lower sidebands at the intermediate frequency are produced by the upper and lower sidebands of the signal. These versions of spectrum conversion have no principal difference. In the case of a frequency-modulated signal the absolute value of frequency deviation is retained after frequency conversion. This means that the signal spectrum also remains unchanged.

“Single-grid” frequency converters employing triodes are most frequently used in wideband ultrashort-wave receivers since multigrid valves produce at these frequencies a high level of internal noise and capacitive couplings between the electrodes become detrimental to receiver operation. Diodes are also used as non-linear mixers. A schematic diagram of a single-grid frequency converter provided with a separate local oscillator employing triodes is illustrated in Fig. 12-31a. Fig. 12-31b shows a frequency converter employing a diode. The processes occurring in single-grid and diode frequency converters were discussed in Chapter XI.

For the sake of simplicity, in Fig. 12-32, the conversion processes of an amplitude-modulated signal in a superhetero-

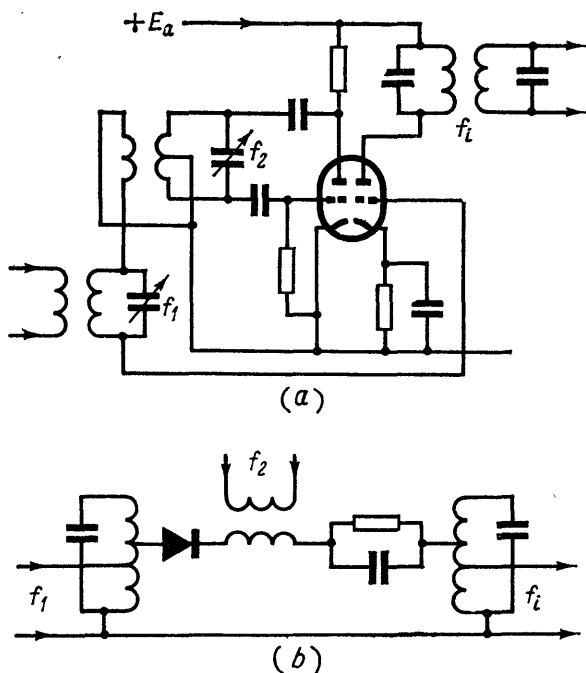


Fig. 12-31. Frequency converters used in ultrashort-wave receiver
(a) converter employing a triode; (b) converter employing a diode

dyne receiver are shown in the course of time. A diode or triode operating on the principle of anode current cutoff is used here as a non-linear element. It is important to understand why signal modulation is preserved after converting a signal to the intermediate frequency: the greater the amplitude of the incoming signal the deeper are its beats with the LO signal at the mixer input and accordingly the greater the IF voltage. A similar interaction between the two signals occurs in electron flow of the mixer valve in the case of two-grid frequency conversion.

Let us now consider transistor frequency converters. One of the possible converter circuits consisting of a mixer and a local oscillator employing the *P-N-P* transistors is given in Fig. 12-33a.

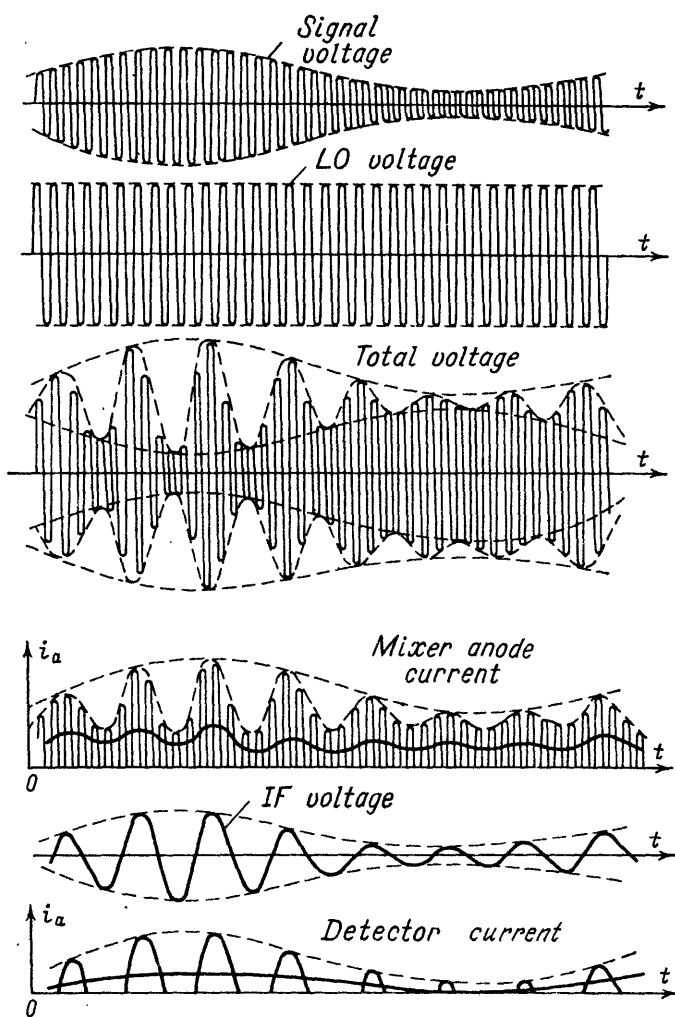


Fig. 12-32. Processes in superheterodyne receiver

Signal voltage u_s fed from an antenna or a preceding amplifier stage is applied to inductor L_1 which excites oscillations in the L_2C_1 circuit tuned to the signal frequency. The tuned circuit is partially coupled with the base of the transistor-mixer, which is expedient because of rather low input impedance of the common-emitter transistor. The circuit consists

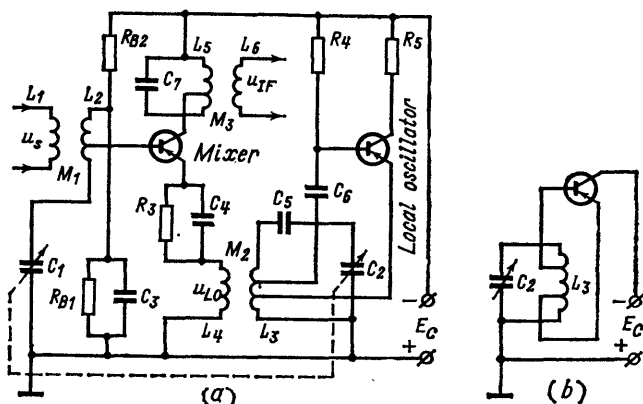


Fig. 12-33. Frequency converter employing transistor
(a) schematic diagram; (b) simplified circuit diagram of a local oscillator

ting of resistors R_{B1} and R_{B2} determines the initial voltage across the base with respect to the common emitter.

Local-oscillator voltage u_{LO} is induced in inductor L_4 coupled to the LO circuit and connected into the emitter circuit. Resistor R_3 is connected into the same circuit to provide temperature compensation of the mixer operating conditions. An IF tuned circuit, L_5C_7 , is inserted into the collector circuit. The collector is connected to some part of inductor L_5 to match the transistor output impedance to the resonance impedance of the tuned circuit. Inductor L_6 couples the IF circuit to the succeeding stage. Incidentally, single-circuit IF stages are quite sufficient for amateur transistor receivers. Needless to say, that a two-circuit filter will provide higher adjacent-channel selectivity but the selection of proper couplings between the two circuits and those between each tuned circuit and the corresponding transistor circuits

presents some difficulties for a radio amateur when adjusting the receiver.

The local oscillator employs a separate common-collector transistor to which the tuned circuit, L_3C_2 , is connected in accordance with a three-point configuration (Fig. 12-33b). Resistor R_5 in the collector circuit prevents parasitic oscillations and protects the collector from damage when accidental current pulses arise in the base circuit. The function of resistor R_4 is to set the required bias voltage for the base, and capacitor C_6 serves for blocking. Finally, capacitor C_5 is included for tracking the signal and local-oscillator circuits at two points of the tuning range.

Physically, the frequency conversion process in a transistor is usually described as a *diode conversion in the transistor input circuit* which is possible owing to a non-linear base current characteristic (see, for instance, Fig. 8-51b) and succeeding IF amplification in the collector circuit. But such a description is simplified since the currents of signal and LO frequencies are also available in the collector circuit and nonlinearities of collector characteristics may also participate in frequency conversion.

Moreover, a secondary component of the signal frequency ($f_{s2} = f_{LO} - f_{IF}$) is produced in the non-linear capacitance C_{12} due to the interaction between the LO and IF oscillations. This effect is known as inverse frequency conversion which may result in unstable operation as the signal-frequency current is fed back to the input circuit through capacitance C_{12} . Inverse frequency conversion is inevitable also in diode mixers but it is absent in two-grid valve converters.

A transistor frequency converter may be designed (for the case of a small signal) with the use of the well-known y -parameters which determine input signal current and output IF current in the terms of signal voltage at the input and IF voltage at the output. Linear equations can be applied since formula (12-22) gives the direct proportionality between the oscillation amplitudes of the signal and intermediate frequency. In particular, the transistor conversion transconductance S_C is assumed to be three times less as compared with transconductance S of the same transistor in the amplification mode. This means that the transistor frequency converter is designed as an *amplifier with deteriorated parameters*.

In conclusion it may be noted that at centimeter waves the diode converter (see Fig. 12-31*b*) which is the most simple among frequency converters takes the form of a specific waveguide similar to that shown in Fig. 12-34. A crystal diode and an LO radiating stub are mounted inside the waveguide section (usually of a rectangular cross section) as illustrated in the figure. Signal waves from an antenna system or microwave amplifier enter the left end of the waveguide.

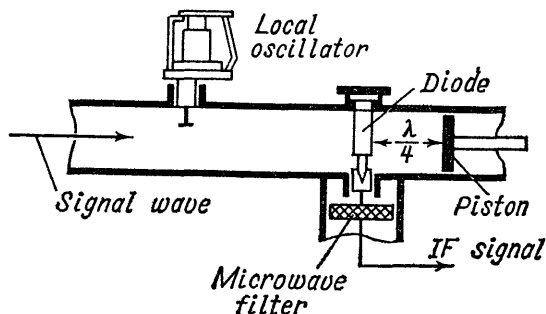


Fig. 12-34. Construction of centimetre-band diode frequency converter

The diode is placed at a quarter-wave distance from the metal piston which closes the right end of the waveguide in order to obtain the maximum signal field at the diode position. In this case, the waves reflecting from the piston are added in phase with the forward waves precisely at the diode position, thus providing maximum energy for frequency conversion. The piston position may be adjusted. The diode position is adjusted relative to the waveguide axis in the process of mixer fabrication, so that to provide the required resistance inserted into the waveguide by the diode.

In the example shown a klystron is used as a local oscillator. Its coupling to the mixer may be adjusted by changing the depth of the stub in the waveguide. A DC microammeter connected into the diode circuit may serve as a tuning indicator. The IF signal is brought out of the waveguide through a cable fitted with a quarter-wave "shut-off" cylinder.

At microwaves diode conversion is employed owing to the low noise level and low input capacitance of the point-contact diode (see Fig. 8-45) although such a frequency converter does not provide signal amplification.

12-7. RECEIVER DETECTOR STAGES

We have agreed earlier to consider detection as the process of conversion of a radio signal into a primary (control) signal. Thus, detection is one of the most important processes performed during radio reception.

Let us consider the detection of an AM radiotelephone signal, which is easy to describe. We are familiar with the

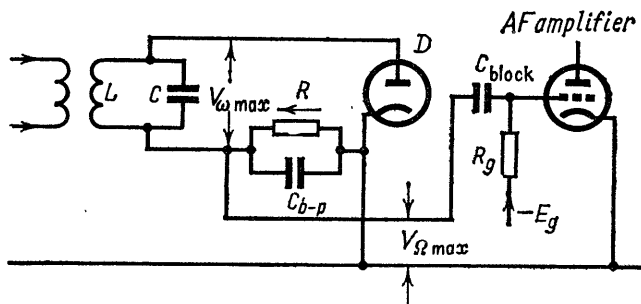


Fig. 12-35. Vacuum diode detector

physical concept of detection of such a signal in a simple receiver (see Fig. 12-2). Detection of an AM radio signal is the process of recovering the primary (modulating) signal from the modulated signal. The detector must be a non-linear device since the radio signal has only the carrier and sidebands, but we want to obtain the signal of audio frequency which is the difference between the carrier and side frequencies. This means that the detection process is similar to the frequency conversion process in a non-linear mixer except that in the detector the carrier itself serves as the LO signal.

The diode (valve or crystal) is the main non-linear device used for detection. The detector is followed by AF amplifier stages. The AF amplifier is controlled by the audio-frequency voltage fed from the detector stage. This voltage is produced by rectified current across a load resistor connected in the detector circuit. A nonwire resistor (hundreds of thousands of ohms for a valve diode) by-passed for radio frequency by a low-capacitance capacitor is generally used as a load.

Thus, we have come to a schematic diagram of a diode detector which is shown in Fig. 12-35. In this diagram the

LC circuit is the final circuit of the RF amplifier (HF or more often IF amplifier). Radio-signal voltage $V_{\omega \max}$ is taken from the tuned circuit to the detector input and due to the presence of a by-pass capacitor C_{b-p} it is fully applied to the anode-cathode terminals of diode D . On one hand, it is desirable to take a sufficiently high capacitance C_{b-p} so that its reactance at high-frequencies should be low. On the other hand, however, it is inadmissible since the capacitor would then appreciably shunt the load resistor R at audio frequencies. It is common practice, therefore, to use capacitors C_{b-p} of the order of hundreds of picofarads when R is measured with several hundreds of kilohms in broadcast bands.

A modulated signal applied to the diode produces a pulsating rectified current (see Fig. 12-2) consisting of a direct component and an audio-frequency component. This means that direct and audio-frequency voltages simultaneously appear across the resistor. The direct voltage across the load resistor has its "plus" and "minus" at its right and left ends, respectively (in compliance with the direction of rectified current shown by the arrow). The presence of DC voltage across the control grid of the succeeding valve should be avoided (since it may displace an operating point of the amplifier to a non-linear portion of its characteristic curve). Therefore, a blocking capacitor C_{block} (0.01 μ F or more) presenting a low reactance for audio-frequency current is connected before the amplifier valve grid so that only an audio-frequency voltage $V_{\Omega \max}$ is applied to it. Normal DC bias E_g is supplied to the amplifier valve grid through grid-leak R_g (hundreds of kilohms, with $R_g > R$).

The DC rectified voltage across resistor R proves to be applied between the cathode and anode of the detector diode. This means that a signal will shift the diode operating point somewhat to the left, with the shift being the greater the higher the resistance R . In other words, the diode conducts only during a fraction of a half-cycle. Under these conditions rectified voltage amounts to about 0.9 $V_{\omega \max}$. Consequently, in the presence of signal modulation the useful amplitude of audio-frequency voltage will be

$$V_{\Omega \max} \approx 0.9 m V_{\omega \max} \quad (12-23)$$

where m is the modulation factor of the signal being received.

For instance, if the signal voltage at the detector input $V_{\omega \max} = 4$ V and the modulation factor $m = 25\%$, then the audio-frequency voltage $V_{\Omega \max} = 0.9 m V_{\omega \max} = 0.25 \times 0.9 \times 4 = 0.9$ V.

The relationships indicated are true at sufficiently high amplitudes of radio signal (over 1 V). RF and IF amplifiers are generally designed to provide such output amplitudes.

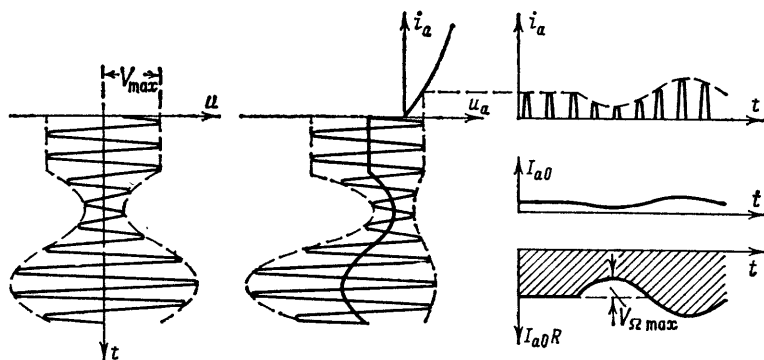


Fig. 12-36. Processes in diode detector

Fig. 12-36 illustrates the processes occurring in a diode detector for the case of large signals. It might be well to point out that negative bias across the diode varies automatically with the changing amplitude of the RF signal so that the cutoff remains unchanged. As a whole, the circuit diagram of the diode detector is very simple but the processes occurring in it are sufficiently complicated.

We now consider in more detail the processes occurring in a valve diode detector under steady-state conditions when an unmodulated signal of large amplitude is applied to it. For this purpose, we display an enlarged view of several oscillation cycles from the preceding figure (Fig. 12-37). Assume that the diode characteristic portion covered by the signal voltage is a straight line with the slope S in the region of positive anode voltages. In the negative region the slope is zero. Such a characteristic curve is known as a piecewise linear one. It is suitable for large signals when a low knee of the actual characteristic curve is of no importance.

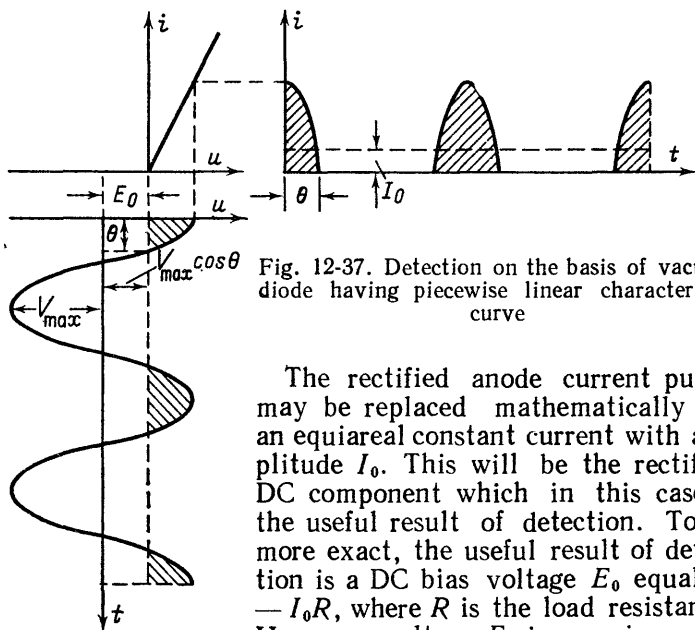


Fig. 12-37. Detection on the basis of vacuum diode having piecewise linear characteristic curve

The rectified anode current pulses may be replaced mathematically by an equiareal constant current with amplitude I_0 . This will be the rectified DC component which in this case is the useful result of detection. To be more exact, the useful result of detection is a DC bias voltage E_0 equal to $-I_0 R$, where R is the load resistance.

However, voltage E_0 is a cosine wave

for an angle of cutoff θ , i.e., $E_0 = V_{\max} \cos \theta$, where $V_{\max}$ is the signal amplitude at the detector input. This means that detector performance may be estimated by the voltage gain defined as

$$K_D = \frac{E_0}{V_{\max}} = \cos \theta \quad (12-24)$$

A remarkable property of a diode having a piecewise linear characteristic curve is in that the *angle of cutoff* does not depend upon the signal amplitude, namely, bias voltage rises with the increasing amplitude in such a way that θ remains constant. The angle of cutoff depends upon transconductance S and load resistance R . For instance, if the product (dimensionless) $SR = 10$, then $\cos \theta \approx 0.65$. If $SR = 100$, then $\cos \theta \approx 0.9$; if $SR = 1000$, then $\cos \theta \approx 0.99$. Thus, the load resistance of several hundreds of kilohms is sufficient to obtain $\cos \theta \approx 0.9$ at normal values of transconductance as indicated above.

When an AM signal is detected, the rectified voltage will change proportionally to the signal amplitude, hence, the audio-frequency voltage which in this case is the useful result of detection will have amplitude

$$V_{\Omega \max} = E_0 m \cos \theta$$

Assume that for a diode $S=2$ mS. If resistance $R=2 \times 10^5 \Omega$ is chosen, then $SR=400$ and we can deliberately take $\cos \theta \approx 0.9$.

Thus, the voltage gain $K_D = \cos \theta$ serves as an important parameter for estimating the diode detector performance. Input impedance is another important parameter which makes it possible to calculate the signal power being consumed by a detector stage, $P_{in} = V_{\max}^2 / 2R_{in}$. Since diode losses are low this power is almost fully developed across resistor R under the action of rectified voltage E_0 , hence $P_{in} \approx E_0^2 / R$. Equating the right-hand sides of the last two equations and setting $\cos \theta \approx 1$, and therefore $E_0 \approx V_{\max}$, we find

$$R_{in} \approx \frac{R}{2} \quad (12-25)$$

If, for instance, the signal produces $V_{\max} = 1$ V and resistance $R = 500$ k Ω , then

$$P_{in} = \frac{V_{\max}^2}{2R_{in}} = \frac{1}{2 \times 2.5 \times 10^5} = \frac{0.2}{10^5} \text{ W (2 } \mu\text{W)}$$

If the receiver sensitivity (input power) is known the required amplification from the input to the detector can be also calculated. For instance, if sensitivity must be 2×10^{-13} W, then in our example $K_p = 2 \times 10^{-6} / 2 \times 10^{-13} = 10^7$ (or 70 dB) will be required. Obviously, the higher the input impedance the better the detector. The value of the load resistance, however, cannot be taken more than hundreds of kilohms since the transients, not discussed here, will distort the recovered primary signal due to a too slow discharge of capacitor C_{b-p} through resistor R .

The distortions of a primary signal waveform can be easily illustrated on the example of the detection of RF pulses. In television receivers the detector should convert an

RF pulse to a video pulse. In this case, the function and schematic diagram of the detector remain the same. In fact, at its output the detector stage must produce a voltage pulse reproducing the envelope of RF pulse being received. A specific feature of pulse detection consists in the selection of circuit parameters. If one keeps the same value for load resistance R and by-pass capacitor C_{b-p} as in the case of radio telephone signal detection, the video pulse will be distorted by transients occurring in the load, namely, the capacitor will gradually charge through the diode and then discharge through the resistor after the RF pulse ends, with the time constants

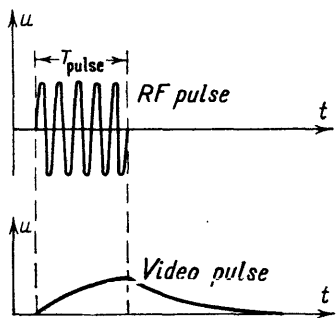


Fig. 12-38. Waveform distortions during pulse detection

being comparable with pulse duration T_{pulse} as shown in Fig. 12-38. One should reduce the value of load resistance R (down to thousands and even hundreds of ohms) and take a low capacitance C_{b-p} in order to cut down the duration of transients (the charge and discharge times of capacitor C_{b-p}). In this case, voltage gain will be much less than 0.9. Input impedance is also reduced.

Diode detection may be accomplished by employing crystal diodes. Crystal diode detectors are extensively used in transistor receivers. Operating principles of such a detector, however, differ significantly from those of the valve diode detector.

The crystal diode operates at low RF voltages (about 0.1-0.3 V) and has a comparatively low load resistance R (Fig. 12-39a). The fact is that a transistor amplifier stage following the detector possesses a low input impedance and, therefore, the value of R must not exceed several kilohms since the two resistances are connected in parallel for alternating current. Naturally, the voltage gain which we assumed equal to 0.9 for a valve detector at large signals will range approximately from 0.1 to 0.3 for crystal diodes (at an RF voltage of about 0.1 V). The input impedance of a diode circuit (at indicated voltages) is calculated approximately from

the formula

$$R_{in}^r = \frac{R}{2K_D}$$

where K_D is the voltage gain.

Thus, for instance, when the AC load resistance $R=1\text{ k}\Omega$ and voltage gain $K_D=0.3$, then $R_{in}\approx 1600\ \Omega$. If signal voltage $V_{\max}=0.1\text{ V}$, then the power consumed by the detector will be

$$P_{in} = \frac{V_{\max}^2}{2R_{in}} = \frac{10^{-2}}{2 \times 1600} = 3\ \mu\text{W}$$

(i.e., has the same order as for a valve diode).

A disadvantage of this simple detector circuit employing a crystal diode is a poor filtering of HF voltage which upon being additionally amplified in the AF stages may feed back

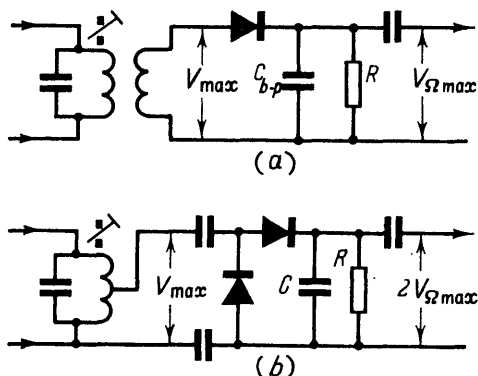


Fig. 12-39. Schematic diagrams of crystal-diode detectors
 (a) ordinary detector; (b) full-wave detector

to the RF (or IF) input, thus causing parasitic oscillations in the receiver. To by-pass resistor R for HF current, capacitance C_{b-p} should be chosen so high that the currents of upper audio frequencies can easily pass through it (causing deterioration of the receiver frequency response).

Somewhat more perfect is a full-wave detection circuit shown in Fig. 12-39b. In this circuit the positive half-cycle of input voltage produces a current pulse through the upper

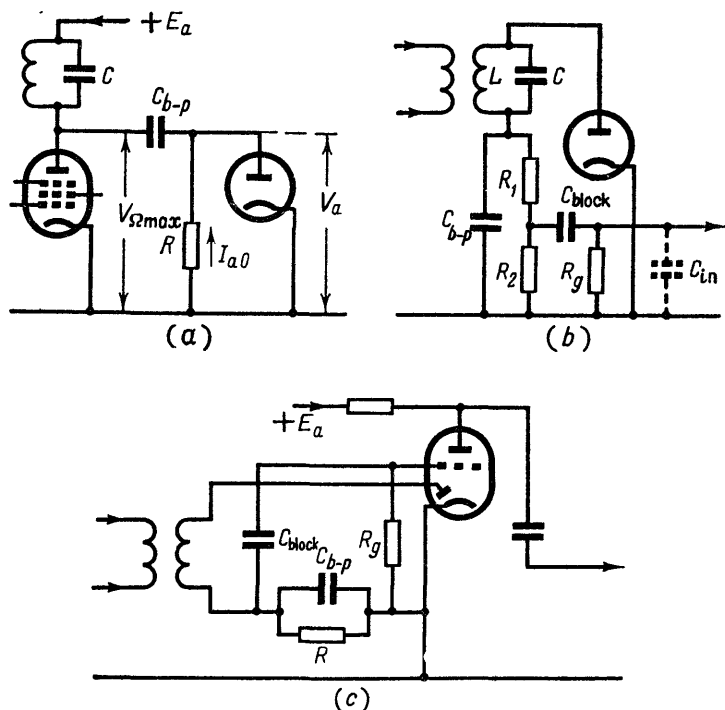


Fig. 12-40. Schematic diagrams of vacuum-diode detectors

(a) detector having a blocking capacitor at its input; (b) detector having an improved HF filtration; (c) detector employing a diode-triode valve

diode and resistor R and the negative half-cycle gives a current pulse in the upper and lower diodes and resistor R , both in the same direction. The circuit provides a doubled rectified voltage (at the given $V_{\max}$) and ensures improved filtering. Input impedance, however, is twice reduced since both half-cycles of the signal are used.

Let us now consider several circuit versions of valve and semiconductor detectors.

If an RF circuit, LC , is coupled directly to the anode circuit of a preceding stage, the diode should be protected against the action of the positive anode supply voltage. In this case, use is made of the schematic diagram shown in

Fig. 12-40a. In this circuit blocking capacitor C_{block} serves as an isolation capacitor and load resistor R is connected in parallel with the diode.

In principle, such a circuit is worse than the basic one (Fig. 12-35) as here an RF voltage is present across load resistor R , i.e., RF filtering is completely absent. In turn, the presence of RF current in the audio-frequency channel may cause receiver parasitic oscillations due to the action of the output on the RF (or IF) input through accidental feedback circuits.

In addition, the parallel connection of resistor R reduces the detector input impedance down to the value of $R/3$ (instead of $R/2$).

In Fig. 12-40b the schematic diagram of a detector with improved RF filtering is shown. Here, useful voltage is taken from some portion (R_2) of the load resistance $R_1 + R_2$. This reduces somewhat audio-frequency voltage. However, if we take into account that resistor R_2 is shunted for high frequency by input capacitance C_{in} of the succeeding stage and that resistor R_1 is not shunted by this capacitance, it becomes clear that the RF voltage will drop primarily across resistor R_1 not acting at the detector output.

Figure 12-40c shows a combination of a diode detector and a resistance-coupled AF amplifier employing a diode-triode valve.

Envelope detection may also be accomplished with the aid of amplifier valves or transistors. Moreover, any circuit of an amplifier device possessing a non-linear characteristic may be used for detection. From this viewpoint, circuits employing amplifier valves may accomplish grid, anode detection or detection with the aid of cathode current. At the present time grid and anode detection is hardly ever used, whereas cathode detection (in cathode follower) illustrated in Fig. 12-41a may still find its application. In case of cathode detection AM voltage is applied between the grid and the cathode of the triode through capacitor C_k having a capacitance of 100-200 pF. A 30-k Ω resistor R_k is connected in parallel with the capacitor. The bias voltage produced across resistor R_k by cathode (anode) direct current shifts the initial operating point to a lower knee of the anode characteristic curve. Detection occurs owing to the non-linearity of this

portion. An audio-frequency component arising in the composition of the cathode (anode) current passes also through resistor R_k producing across it the useful voltage which is taken to volume-control resistor R through blocking capacitor C_{block} having a large capacitance.

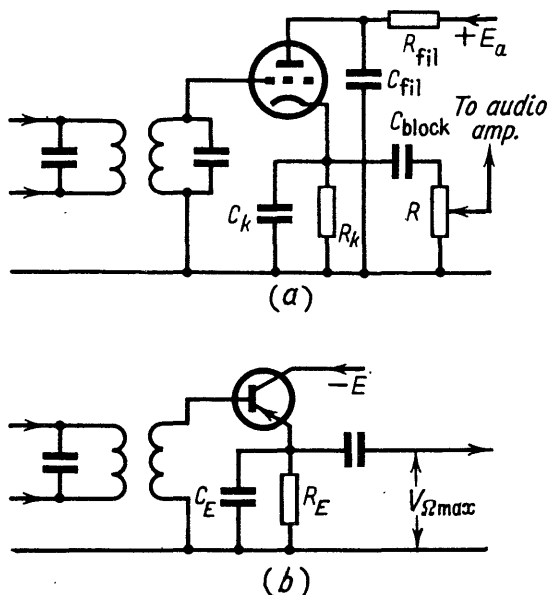


Fig. 12-41. Schematic diagrams of detectors employing amplifier devices
(a) cathode-follower detector; (b) emitter-follower detector

An advantage of the cathode detector as compared with the diode one is its higher input impedance. As with any cathode follower, this circuit gives no amplification.

Similarly, to provide a high input impedance for the detector in a transistor receiver, it is advantageous to use a transistor having a load resistor in its emitter circuit, i.e., a common-collector circuit or an emitter follower as shown in Fig. 12-41b.

Let us now turn to the detection of a *frequency-modulated radiotelephone signal*. As known from Chapter XI, the envelope detector is unusable for this purpose since an FM signal

has a fixed amplitude and, therefore, only direct current is produced when it is applied to the input of an ordinary (for instance, diode) detector. FM signal amplitude may change in the result of external interference and internal noise, but these are undesirable variations.

Thus, we can state that an FM receiver should meet the following requirements: first, to eliminate amplitude variations of FM signal with the aid of a limiter; second, to convert

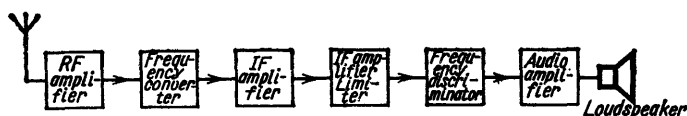


Fig. 12-42. Block diagram of FM receiver

the FM signal to the AM one; third, to perform ordinary envelope detection of the latter. Sometimes, the FM-AM conversion and detection of a converted signal are accomplished in a single unit. An example of such a combined unit known as a discriminator has been given in Chapter XI. Other schemes for detection of FM signals are also possible. The name "FM discriminator" is adopted for all the combined units designed for this purpose.

Proceeding from the above, we may represent a block diagram of an FM receiver as shown in Fig. 12-42 which differs somewhat from that of the AM receiver given in Fig. 12-6. The last IF amplifier stage of such a receiver is usually placed in an amplitude limitation mode though a separate limiter can also be used.

However, not only the limiter and converter constitute the characteristic properties of FM reception. The distinguishing feature of frequency modulation is the required bandwidth and, as a consequence, the used wavelength band.

A great modulation index (see Chapter XI) is required in FM broadcasting since only under this condition can complete interference elimination be obtained owing to the "predominance" of useful frequency modulation over the parasitic one produced by interference. A great modulation index is associated with an appreciable frequency deviation. In broadcasting the maximum deviation (corresponding to the loudest sound) amounts to about 100 kHz. In turn, the required band-

width, as we know, is equal to a doubled value of the maximum deviation. Consequently, the FM receiver should have the bandwidth

$$B = 2\Delta f_{\max} \approx 200 \text{ kHz}$$

By assuming an IF-circuit quality factor $Q_{eq} = 50$ and considering that the bandwidth $2\Delta f_{\max} \approx f_i / Q_{eq}$, we find a suitable value of intermediate frequency $f_i \approx 2\Delta f_{\max} Q_{eq} = 200 \times 50 = 10,000 \text{ kHz}$ ($\lambda = 30 \text{ m}$). In actual FM receivers, the intermediate frequency is chosen approximately in this region, i.e., from 8.4 to 10.7 MHz. Thus, radio frequency in this case should certainly lie within the ultrashort wave band so that its portions may alternate with those of a television broadcast band.

We may now state the major features differing the FM broadcast receiver from the AM receiver, namely, the presence of an FM discriminator and an amplitude limiter, a wide bandwidth and, what is most important, an ultrashort wave band. It is clear that high-class multi-purpose receivers are capable of receiving not only AM but FM signals as well, with the change-over to FM reception being accomplished by switching to the corresponding RF and IF circuits.

Let us consider some typical varieties of limiter and discriminator circuits used in modern broadcast receivers.

In Chapter XI we described the operating principles of the simplest limiter circuits employing diodes and triodes. Amplitude limitation in FM receivers is generally accomplished in the last IF amplifier stage employing a pentode rather than in a separate stage. For this purpose resistor R_{g1} (about 100 k Ω) by-passed with capacitor C_{g1} (about 50 pF) is connected into the control grid circuit and a reduced voltage is applied to the screen grid (through a 150-k Ω resistor R_{g2}) and to the anode (Fig. 12-43a). As the signal increases, a threshold is set up by the current arising in the first grid and because of the presence of the upper knee in the anode characteristic curve. It is desirable that the gain of this stage at low signal voltages (below the threshold) be approximately of the same value as in ordinary IF amplification mode. The amplitude characteristic of such a stage, i.e., the relationship between its output voltage $V_{out \max}$ and signal voltage $V_{g1 \max}$ (Fig. 12-43b) is generally taken so that the performance of the

limiter may be estimated. As seen from the characteristic curve, the limiting threshold corresponds to an input amplitude of about 1 V and that up to this threshold the stage gain $K_0 \approx 20$. Limiting of the RF signal subjected to a spurious amplitude modulation is illustrated in Fig. 12-43c.

The output circuit of the limiter simultaneously serves as one of the discriminator circuits described earlier (see

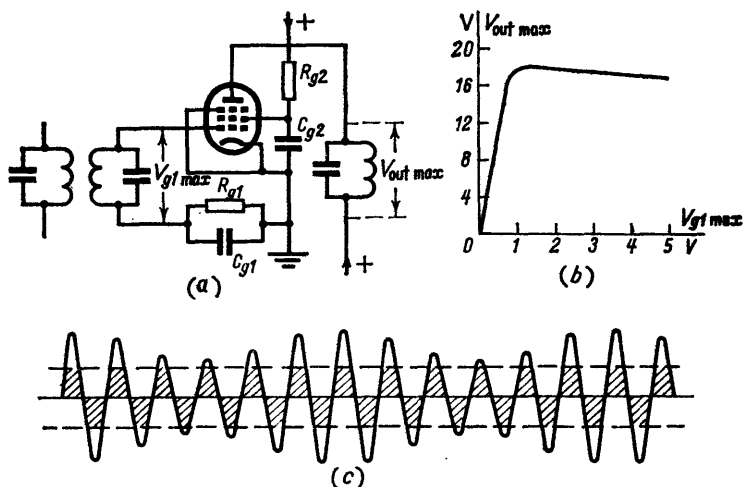


Fig. 12-43. Limiter stage

(a) schematic diagram; (b) amplitude characteristic curve; (c) illustration of limiting process

Chapter XI). In comparatively simple receivers, however, the discriminator often performs yet the function of an amplitude limiter so that the last IF stage can operate as an amplifier only. The discriminator functioning as a limiter, modulation converter and amplitude detector is called a *ratio detector*. One of its possible schemes is illustrated in Fig. 12-44.

The major difference between the ratio detector and the discriminator described in Chapter XI is in that diodes D_1 and D_2 are connected in series rather than back-to-back and their load resistors R_1 and R_2 are shunted by capacitor C_4 (about a dozen of microfarads). A centre lead is provided between capacitors C_2 and C_3 having a low capacitance, such

that their reactances are great for audio frequency. Output audio-frequency voltage $V_{\Omega \max}$ is taken between this lead and a grounded centre lead provided between resistors R_1 and R_2 , and then is filtered by RF filter R_3C_5 .

Assume that the incoming signal has spurious amplitude modulation. The voltage detected by diodes and developed across resistors R_1R_2 would pulsate at the spurious modulation frequency. The capacitor C_4 , however, has no time to

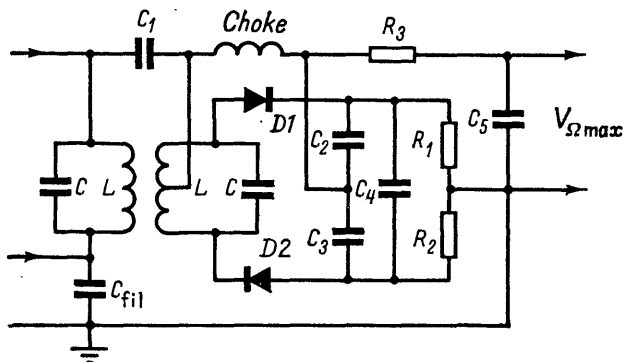


Fig. 12-44. Ratio detector

charge and discharge with the audio-frequency rate since the product $C_4(R_1+R_2)$, i.e., the time constant is great as compared with the audio-frequency cycle, so that voltage across this capacitor remains constant. It is precisely this constancy that is equivalent to the action of an amplitude limiter.

In the presence of frequency modulation, as we know from the description of discriminator operation (see Chapter XI), the voltages across the diodes change, i.e., when voltage across one of them rises, it drops across the other and vice versa. Accordingly, rectified voltages across capacitors C_2 and C_3 also vary with the audio-frequency rate but their sum remains constant and equal to the voltage across capacitor C_4 . Thus, the sum voltage is redistributed between capacitors C_2 and C_3 . In other words, if we take the ratio of the above voltages, V_{C_2}/V_{C_3} , then it is this ratio which changes with the frequency modulation rate and, therefore, the circuit has the name of ratio detector. The voltage taken between the centre

point of capacitors C_2 and C_3 and the grounded centre point of resistors R_1 and R_2 is then applied to the input of the audio-frequency amplifier.

12-8. CONTROLS AND ADJUSTMENTS IN RECEIVERS

Some kinds of controls and adjustments have already been described in preceding chapters. We now discuss in more detail the adjustments used in different classes of receivers. One or the other parameter or receiver characteristic may be adjusted to provide the most favourable conditions for receiving the desired radio station. Receivers are provided with both manual and automatic controls. Moreover, in high-performance receivers both types of control may supplement each other.

Among the major kinds of controls are, of course, *tuning controls*. Tuning is accomplished most often manually. Subbands (i.e., RF circuit inductances) are changed-over with the aid of a rotary or push-button switch. Continuous tuning within a subband is carried out by turning the rotor axle of the gang capacitor.

In comparatively rare instances, continuous tuning is supplemented with push-button tuning to the wavelength which is fixed for each button. In the simplest case the gang capacitor is set in the position corresponding to the required wavelength by pushing the button. In more complicated cases, button pushing results in sudden changes of inductance or capacitance, i.e., the button mechanism serves as a switch for the RF and LO circuits. It will be recalled that three circuits (input circuit, amplifier and LO circuits) require tuning when the receiver has only one RF amplifier stage. In some of the most complicated receivers a small electric motor rota-

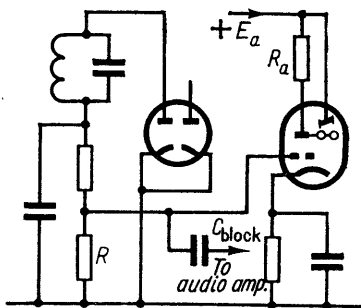


Fig. 12-45. Circuit connection of cathode-ray tuning indicator ("magic eye")

tes the gang capacitor rotor axle when started by a push-button. The motor stops by breaking the supply network at a preset angle of turn.

Most valve receivers are provided with a *cathode-ray tuning indicator* ("magic eye") which facilitates fine circuit tuning to resonance with the incoming signal. Such a tuning indicator is a combination of a triode and a special diode, which have a common cathode and are enclosed in one envelope. A screen (target) made in the form of a cup serves as a diode anode. The inner (concave) surface of the target is coated with a fluorescent material which fluoresces with greenish glow under electron bombardment. The cup contains a metal rod running from the triode anode. The valve is secured in place at the back of the receiver panel through an opening which allows the operator to observe the fluorescent screen.

The tuning indicator is connected after the detector (Fig. 12-45). Direct voltage across the load resistor R isolated from the amplifier input with capacitor C_{block} is fed to the indicator grid without a blocking capacitor. A high-resistance resistor R_a is connected between the triode anode and the fluorescent target. When signal is absent and no negative bias is applied to the indicator target, appreciable current flows in the triode portion through resistor R_a , therefore, the anode potential becomes below that of the target. In this case the metal rod repels electrons and a wide dark sector (a wedge-shaped shadow) may be observed on the screen. When a signal is received, the triode becomes cut off, the potential of its anode and metal rod rises, and the dark sector narrows, which is an indication of tuning to the radio station.

It is very convenient to tune the receiver against the indicator with volume reduced. In this case while tuning over the receiver scale, one cannot hear undesired stations. When the receiver has been tuned to the desired station, volume should be increased. Of course, indicator failure does not disturb the normal operation of the receiver.

Gain controls are of great importance for obtaining the normal conditions of reception. As we know, high gain providing high sensitivity of the receiver is required for reception of signals from distant or low-power transmitters. However, high gain is detrimental when receiving radio waves of high field intensity since the receiver stages will be overloaded and

the sound produced by the loudspeaker will be very loud and distorted.

In comparatively simple receivers the gain of HF and IF stages may be controlled manually. When it is possible to provide a negative bias source— E_g in the receiver supply system, its value across the valve control grids can be controlled manually by potentiometer R as shown in Fig. 12-46 for two IF stages. In adjustable stages it is advisable to use

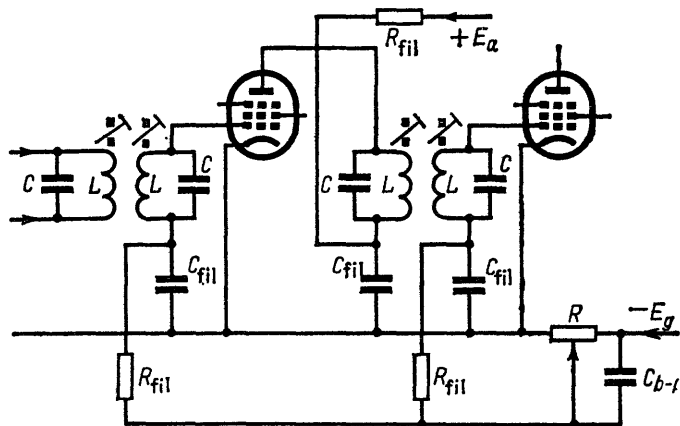


Fig. 12-46. IF manual gain control

the pentodes whose characteristics have an extended tail similar to those shown in Fig. 8-37. In the case of a large signal, for which a high negative bias is set by the gain control, the operating portion of the characteristic will be linear so that the envelope of an AM signal will not be distorted. Non-linear distortions of a modulated signal envelope are possible when pentodes with short characteristics are employed. These distortions appear after the detector as higher harmonics of the audio signal. It is precisely for this purpose that long-tail characteristic pentodes were designed.

When a separate source is used to supply negative bias voltage to the grids of the valves being controlled, it is not expedient to connect cathode resistors in their circuit since anode current will decrease with the increasing bias voltage

and, therefore, the voltage drop across the cathode resistor will also decrease, thus weakening the effect of gain control.

A manual gain control should be used when transferring from the reception of large signals (from power and nearby radio station) to the reception of weak signals (or vice versa), thus providing in each case an approximately constant amplitude at the detector input corresponding to the normal detection mode. This may be accomplished manually but it is much more difficult to follow variations in the field intensity of a single radio station with the aid of the manual gain control. At short waves and during night at medium waves, field intensity may suddenly change owing to fading. In automobile receivers incoming-signal field intensity may vary severely when moving behind obstructions (reinforced concrete buildings, bridges, hills, etc.). In such cases it is much more convenient to provide automatic gain control (AGC) for the IF (in some instances, for HF) amplifier stages.

How many stages should be covered by the AGC circuit? It is common practice to specify how many times less the voltage variations must be at the detector input than those of signal field intensity. Proceeding from this, the required variation in the total predetection gain $K_{\max}/K_{\min}$ can be calculated. Gain is proportional to the transconductance and if stages are identical, then $K_{\max}/K_{\min} = (S_{\max}/S_{\min})^N$, where N is the number of the stages being controlled and $S_{\max}$ and $S_{\min}$ are the attainable limits of transconductance variations.

The characteristic feature of AGC in comparison with the manual gain control is the generation of a negative bias voltage by means of radio signal rectification. For this purpose, we could use the rectified voltage DC component across the load resistor of the diode detector, i. e., the same DC component which we used to control the tuning indicator as shown in Fig. 12-45. Such a simple AGC circuit is applied in practice. The disadvantage of this circuit is in that the detector produces a negative rectified voltage even in the presence of weak signals, thus deteriorating the receiver sensitivity.

A delayed automatic gain control employing a separate diode is much better in operation. It is precisely for this purpose that double diodes or double diodes in combination with triodes or pentodes are produced. The schematic diag-

ram used for generating a negative bias voltage from a separate diode for a delayed AGC is illustrated in Fig. 12-47. The delay is obtained owing to a DC positive voltage applied to the cathode of the right-hand (in the diagram) diode (AGC diode) from resistor R_h which is fed with the cathode current of an AF amplifier stage. Consequently, a DC negative voltage is applied to the diode anode. No bias voltage is applied

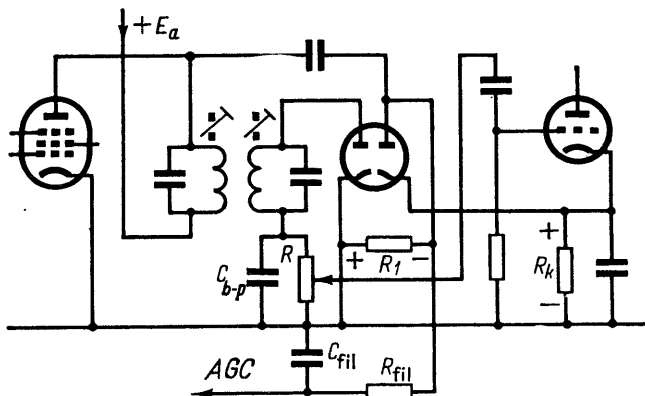


Fig. 12-47. Voltage source for delayed automatic gain control

to the grids of the valves being controlled until the signal amplitude exceeds the delay voltage. When the signal exceeds the delay voltage, the rectified voltage fed to the control grids of the valves being controlled is produced across resistor R_1 connected in the AGC diode circuit. This means that a further increase in the signal amplitude leads to a bias rise across the control grids of the valves being controlled and, consequently, to a decrease in the gain of HF and IF stages. In this way, variations in voltage applied to the detector input are automatically and sharply limited as compared with those in signal field intensity. The low-frequency filter $C_{fil}R_{fil}$ is connected into the circuit to avoid the action of AC voltages from the load resistor R_1 of the AGC diode on the valve grids. It should be noted that IF voltage is taken from the 1st band-pass filter circuit in a parallel way (see Fig. 12-40a), whereas the voltage applied to the detector diode is taken from the 2nd bandpass filter circuit in a series way. Owing to such

connections the quality factors of both bandpass filter circuits are retained approximately equal.

Moreover, AGC provides an approximately constant volume since it maintains a relatively constant signal at the detector input. This is not always desirable for a listener since volume should be adjusted to suit the room size, an ambient noise level, the time of a day, etc. Therefore, manual gain control in the AF channel is used along with AGC for

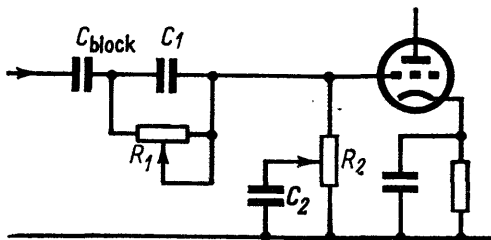


Fig. 12-48. Manual tone controls in valve receiver

HF and IF amplifier stages. An audio-frequency voltage divider connected in the load circuit (potentiometer R shown in Fig. 12-47) of the detector diode may be used as a simple manual *volume control*.

More complex functions are accomplished by manual *tone controls* in the AF amplifier stages. Tone colour depends upon amplification of different tones of the sound-signal spectrum. In principle, the tone colour of the sound being reproduced should be similar to that of the primary signal which modulates the transmitter. If gain is reduced in the lower audio-frequency region, the sound acquires a jingling tone which emphasizes the basic melody. If gain is reduced in the higher audio-frequency region, the sound will be more dull emphasizing the music accompaniment. In turn, if gain is reduced at lower and higher audio frequencies, the receiver frequency response (see Fig. 12-4) narrows from the left and right, thus decreasing the noise level at the expense of deteriorating the tone colour of the incoming signal.

Thus, one may encounter two tone controls, namely, a lower tone control and higher tone control. The simplest example of the circuits for such tone controls at the input of

one of the AF amplifier stages is shown in Fig. 12-48. In this schematic diagram capacitor C_{block} is an ordinary blocking capacitor used to couple the preceding detector or amplifier stage to the succeeding one. Capacitor C_1 is of small capacitance and passes only the currents of higher frequencies. By reducing the resistance of variable resistor R_1 connected in parallel with capacitor C_1 , we permit the lower frequency currents to flow through the circuit. The capacitance of capacitor C_2 is appreciable. If it is fully in parallel with grid resistor R_2 , the amplifier input will be completely shunted for the higher frequency currents. Gain at higher frequencies will rise if the wiper of resistor R_2 is moved down (in the schematic diagram).

12-9. EXAMPLES OF RECEIVER CIRCUITRY

When considering the operating principles of all the units of a receiver, we should include some examples of the circuitry in which these units are assembled, thus making the complete receiver.

A schematic diagram of a superheterodyne receiver employing electron valves is illustrated in Fig. 12-49. For the simplicity of displaying the major parts of the circuitry, it is shown for one band only.

The receiver has the following units: input circuits with an inductively coupled antenna, RF amplifier stage employing a pentode and the inductive coupling between the tuned circuit and anode network, frequency converter employing a heptode provided with a two-circuit IF filter, two IF amplifier stages employing pentodes with the identical filters, detector (left-hand diode of the double diode-triode), automatic gain control for three stages (right-hand diode of the double diode-triode), audio-frequency amplifier stage (triode portion of the double diode-triode) and output stage employing a beam tetrode. All the units of this receiver have been discussed in preceding sections.

We may present here only some typical units of all-wave multivalve superheterodyne receivers whose circuitry is bulky. A schematic diagram illustrating the switching of extended bands in the vicinity of 25-m, 31-m and 41-m wavelengths ("daytime", "evening" and "night" portions of the

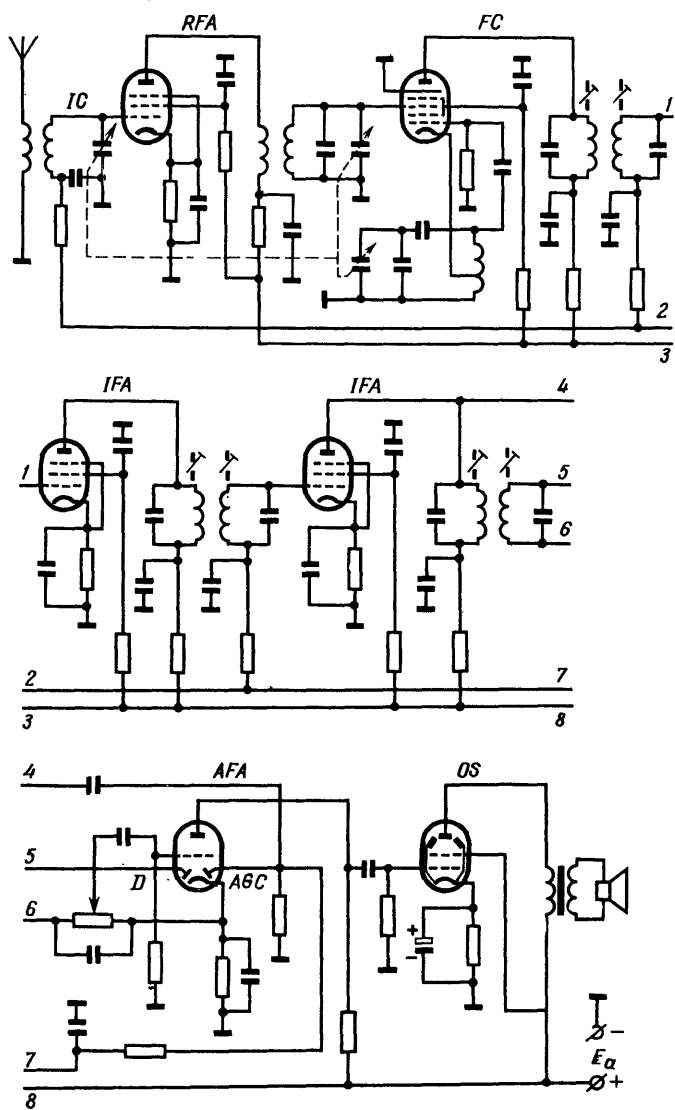


Fig. 12-49. Superheterodyne receiver employing electron valves

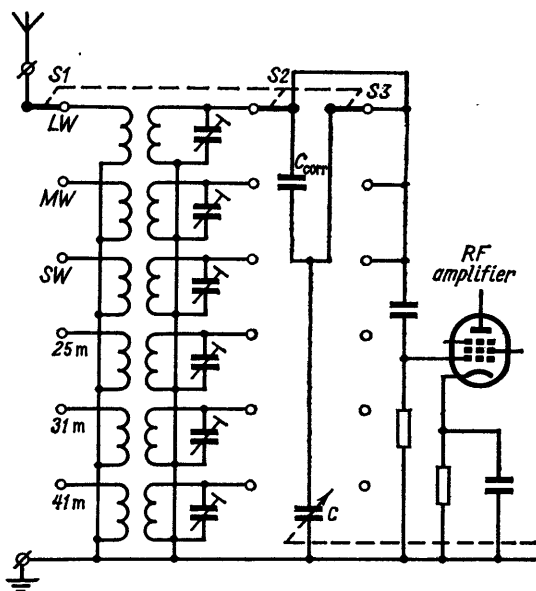


Fig. 12-50. Input-unit band switch

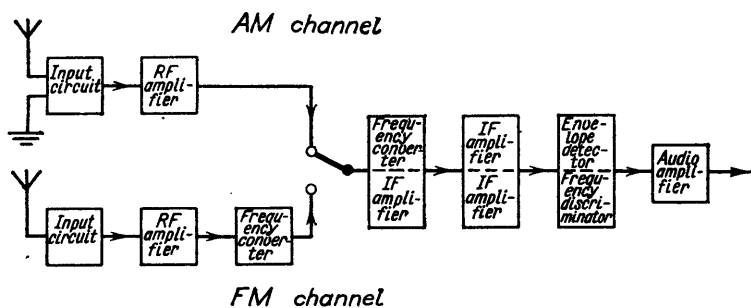


Fig. 12-51. Block diagram of AM-FM broadcast receiver

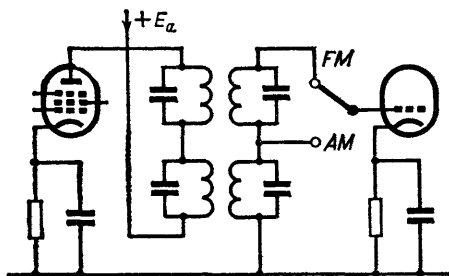


Fig. 12-52. Switching of IF filters

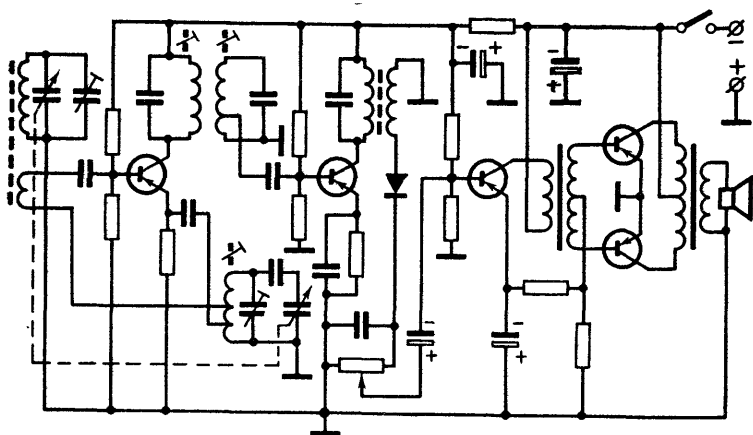


Fig. 12-53. Simplest transistor superheterodyne receiver

short-waveband, see Chapter VII) is illustrated in Fig. 12-50. Switches S_1 , S_2 and S_3 are mechanically coupled to one another. The first two switches actuate the coils of one or another subband, whereas switch S_3 serves to short the correcting capacitor C_{corr} for main subbands and to connect it in series with a tuning capacitor C for extended subbands where capacitor C_{corr} is used to reduce the subband overlap inherent in the tuning capacitor.

A block diagram of a superheterodyne receiver having standard AM signal bands as well as one FM ultrashort-wave

band is presented in Fig. 12-51. The following is remarkable—valves of the IF amplifier are the same but bandpass filters are changed when transferring to ultrashort waves. When operating at ultrashort waves, the IF amplifier has an additional stage which ensures sufficient voltage at the input of the amplitude limiter.

An exemplary schematic diagram showing the switching of IF filters is illustrated in Fig. 12-52. The bandwidth of the AM subband filter may be within 8 to 10 kHz, whereas the FM channel bandwidth amounts from 100 to 200 kHz. Intermediate frequencies are taken about 465 kHz and 8 MHz, respectively.

Figure 12-53 illustrates a schematic diagram of the simplest transistor superheterodyne receiver. The magnetic antenna is part of the input unit. The 1st transistor functions as a frequency converter with a *built-in local oscillator*, i.e., the LO circuit is connected to the same transistor through a three-point common-collector circuit. The receiver provides no amplification at the signal frequency. The 2nd transistor is an IF amplifier with a single tuned circuit connected to the diode detector. The 3rd transistor is an AF driver and the output push-pull stage is formed by the last two transistors. All the elements of the receiver have been discussed in the preceding sections.

We may now expect that this our description of radio receivers and their components will allow the reader to understand the circuitry of broadcast receivers. Needless to say that special receivers which are developed concurrent with broadcast receivers and even more intensively have their own distinctive features. In special receivers the search for the desired radio station and tuning adjustments during reception should be eliminated. This is achieved owing to the high stability of the receiver calibration but, first of all, to the high stability of the local oscillator in the frequency converter, which can be provided, for instance, by means of a quartz-crystal control. Such a control, however, excludes continuous tuning of the local oscillator and, therefore, forces the designers to change significantly the receiver circuitry. IF stages employ bandpass filters with many tuned circuits or electromechanical filters (for instance, filters fitted with quartz-crystal plates as oscillatory systems).

The electroacoustic properties of receivers are of great importance for improving broadcast reception. Separation of upper and lower audio frequencies and their reproduction with different loudspeakers expands the sound frequency band and eliminates the distorting non-linear interaction between components of different frequencies in the amplifier. Great progress has been made in the stereo sound reproduction, i.e., in the setting-up of the same acoustic field which acted on microphones in a concert hall.

REFERENCE DATA

Measurement Units of Physical Quantities

Today, calculations in all the fields of science, engineering and economy are preferably based in the Soviet Union on the International System of Units (SI). Other systems of units and arbitrary units are sometimes also employed in the Soviet Union and other countries.

The basic and additional SI units are:

Name	Unit of measurement	Name	Unit of measurement
Length	metre (m)	Luminous intensity	candela (cd)
Mass	kilogram (kg)	Plane angle	radian (rad)
Time	second (s)	Quantity of matter	mol (mol)
Current	ampere (A)	Space angle	steradian (sr)
		Temperature	kelvin (K)

The derived units of SI are as follows:

Frequency	hertz (Hz)
Force, weight	newton (N)
Pressure, mechanical stress	pascal (Pa)
Work, energy, quantity of heat	joule (J)
Power, heat flux	watt (W)
Quantity of electricity, electric charge	coulomb (C)
Voltage	volt (V)
Capacitance	farad (F)
Resistance	ohm (Ω)
Conductance	siemens (mho, S)
Magnetic flux	weber (Wb)
Magnetic induction	tesla (T)
Inductance	henry (H)
Luminous flux	lumen (lm)
Illumination	lux (lx)

Multiple and fraction prefixes. The following prefixes are written directly before the unit to which they belong:

Factor	Prefix	Factor	Prefix
10^{12}	tera (T)	10^{-2}	centi (c)
10^9	giga (G)	10^{-3}	milli (m)
10^6	mega (M)	10^{-6}	micro (μ)
10^3	kilo (k)	10^{-9}	nano (n)

10^2	hecto (h)	10^{-12}	pico (p)
10^1	deca (da)	10^{-15}	femto (f)
10^{-1}	deci (d)	10^{-18}	atto (a)

Relationships between various units of measurement of physical quantities.

Length. 1 angström = 1 Å = 10^{-1} nm = 10^{-4} μm = 10^{-10} m.

1 micron = 1 μm (formerly denoted as μ) = 10^{-3} mm = 10^{-6} m.

1 statute mile = 1 mi = 5280 ft = 1.609344 km = 1609.34 m.

1 fathom = 6 fath = 2 yd = 1.8288 m.

1 yard = 1 yd = 3 ft = 12 in = 914.4 mm = 0.9144 m.

1 foot = 1 ft = 12 in = 304.8 mm = 0.3048 m.

1 inch = 1 in = 1000 mils = 25.4 mm = 0.0254 m.

1 mil = 10^{-3} in = 25.4 μm = 0.0254 mm = $2.54 \cdot 10^{-5}$ m.

Area. 1 barn = 1 b = 10^{-24} cm² = 10^{-28} m².

1 are = 1 a = 1 dam² = 100 m².

1 hectare = 1 ha = 1 hm² = 100 a = 10^4 m².

1 square mile = 1 sq mi = $2.59 \cdot 10^6$ m².

1 square foot = 1 sq ft = 144 sq in = 929.0 cm² = 0.0929 m².

1 square inch = 1 sq in = 10^6 square mils = 645.2 mm² = $6.452 \cdot 10^{-4}$ m².

1 circular inch, i.e., the area of a circle with a diameter of 1 in = 1 circ in = 0.7854 sq in = 506.71 mm² = $5.0671 \cdot 10^{-4}$ m².

1 circular mil, i.e., the area of a circle with a diameter of 1 mil = 1 circ mil = 0.7854 sq mil = $5.0671 \cdot 10^{-4}$ mm² = $5.0671 \cdot 10^{-10}$ m².

Volume. 1 litre = 1 l = 1.000028 dm³ = $1.000028 \cdot 10^{-3}$ m³.

1 cubic foot = 1 cu ft = 1728 cu in = 28.32 dm³ = 0.02832 m³.

1 cubic inch = 1 cu in = 16.387 cm³ = $1.6387 \cdot 10^{-5}$ m³.

Time. 1 technical year = 8760 h.

1 mean solar year = 8765 h 48 min 46 s.

1 hour = 1 h = 60 min = 3600 s.

1 minute = 1 min = 60 s.

1 second = 1 s = 1 ohm · F = 1 Mohm · μF = $9 \cdot 10^{11}$ ohm · cm = 9×10^9 ohm · m.

Linear velocity. 1 km/h = 27.78 cm/s = 0.2778 m/s.

1 foot per minute = 1 ft/min = 0.508 cm/s = 5.08 mm/s = $5.08 \cdot 10^{-3}$ m/s.

1 foot per second = 1 ft/s = 30.48 cm/s = 304.8 mm/s = 0.3048 m/s.

Frequency, frequency of rotation, angular velocity. 1 hertz = 1 Hz = 1 cycle per second.

1 revolution per minute = 1 rev/min = $\pi/30$ rad/s = 0.1047 rad/s.

1 revolution per second = 1 rev/s = 60 rev/min = 2π rad/s = 6.284 rad/s.

1 angular degree per second = 1°/s = 0.0174533 rad/s.

Mass. 1 gram = 1 g = 10^{-3} kg.

1 metric ton = 1 t = 1 Mg = 10^3 kg.

1 pound avoirdupois = 1 lb = 16 ounces = 256 drams = 7000 grains = 453.5924 g = 0.4535924 kg.

1 ounce = 1 oz = 28.94953 g = 0.02834953 kg.

1 metric carat = 1 ct = 200 mg = 0.2 g = 0.0002 kg.

Force. 1 kilogram-force = 1 kgf = 10^3 gf = 10^{-3} tf = 1 kp = 9.80665×10^5 dyn = 9.80665 N.

1 gram-force = 1 gf = 980.665 dyn = 0.00980665 N.

1 ton-force = 1 tf = 1000 kgf = 9806.65 N.

1 dyne = 1 dyn = $10\mu\text{N} = 10^{-5}\text{N}$.

1 pound-force = 1 lb.-f = 453.5924 gf = 4.44822 N.

Pressure, mechanical stress. 1 pascal = 1 Pa = 1 N/m².

1 bar = 1000 millibars = $10^5\text{ N/m}^2 = 10^5\text{ Pa}$.

1 millibar = 1 mbar = $100\text{ N/m}^2 = 100\text{ Pa}$.

1 physical atmosphere = 760 mm Hg = $1.01325 \cdot 10^5\text{ N/m}^2 = 1.01325 \cdot 10^5\text{ Pa}$.

1 technical atmosphere = 1 kgf/cm² = 98066.5 Pa.

1 dyn/cm² = 0.1 N/cm² = 0.1 Pa.

1 millimetre of mercury column = 1 mm Hg = 133.322 N/m² = 133.322 Pa.

1 millimetre of water column = 1 mm H₂O = 1 kgf/cm² = 9.80665 N/m² = 9.80665 Pa.

1 pound-force per square inch = 1 lb.-f/sq in = 1 psi = 6894.076 N/m² = 6894.076 Pa.

Specific impact strength. 1 kgf.cm/cm² = 1 kgf/cm = 100 kgf/m = 980.665 N/m = 980.665 J/m².

1 pound-force-foot per square inch = 1 lb.-f × ft/sq in = 2.143 kgf/cm = 214.3 kgf/m = 2101.56 N/m.

Dynamic viscosity. 1 poise = 1 P = 0.1 N.s/m² = 0.1 Pa.s.

1 centipoise = 1 cP = 0.01 P = 1 mN.s/m² = 0.001 N.s/m² = 0.001 Pa.s.

Kinematic viscosity. 1 stokes = 1 St = $\frac{\text{P}}{\text{g/cm}^3} = 0.001 \frac{\text{P}}{\text{kg/m}^3} = 10^{-4}\text{m}^2/\text{s}$.

1 centistokes = 1 cSt = $10^{-6}\text{ m}^2/\text{s}$.

Temperature. The ratio between the temperatures in K (absolute temperature), °C and °F is as follows:

$$T, \text{ K} = T, \text{ }^\circ\text{C} + 273.16 = \frac{5}{9} (T, \text{ }^\circ\text{F} - 32) + 273.16;$$

$$T, \text{ }^\circ\text{C} = \frac{5}{9} (T, \text{ }^\circ\text{F} - 32)$$

The temperature difference is expressed in kelvins. When the difference of temperature is compared in K and °F we have:

$$\Delta T, \text{ K} = \frac{5}{9} \Delta T, \text{ }^\circ\text{F}$$

Energy. 1 joule = 1 J = 1 W.s = 1 V.C = 1 N.m.

1 erg = 1 dyn.cm = $0.1\mu\text{J} = 10^{-7}\text{ J}$.

1 kilowatt-hour = 1 kW.h = 3600 kJ = 3.6 MJ = $3.6 \cdot 10^6\text{ J}$.

1 electron-volt = 1 eV = 0.16021 aJ = $1.6021 \cdot 10^{-19}\text{ J}$.

1 kilogram-force-metre = 1 kgf.m = 9.80665 J.

1 pound-force-foot = 1 lb.-f × ft = 1.35582 J.

1 calorie = 1 cal = 4.1868 J.

1 kilocalorie = 1 kcal = 1000 cal = 4.1868 kJ = 4186.8 J.

1 British thermal unit—quantity of heat raising 1 lb of water by 1 °F = 1 B. T. U. = 1.0548 kJ = 1054.8 J.

Power. 1 watt = 1 W = 1 J/s = 1 VA = 1 V²/ohm = 1 N.m/s.

1 kilogram-force-metre per second = 1 kgf.m/s = 9.80665 W.

1 calorie per second = 1 cal/s = 4.1868 W.

- 1 kilocalorie per second = 4.1868 kW = 4186.8 W.
 1 kilocalorie per hour = 1 kcal/h = 1.163 W.
 1 horse power = 1 hp = 735.5 W.
Specific heat capacity. 1 J/(g·K) = 1 kJ/(kg·K).
 1 cal/(g·K) = 1 kcal/(kg·K) = 1 B. T. U./(lb × °F) = 4.1868 J/(g·K) =
 = 4.1868 kJ/(kg·K) = 4.1868 · 10³ J/(kg·K).
 1 erg/(g·K) = 10⁻⁴ J/(kg·K).
Specific heat of phase transition.
 1 cal/g = 1 kcal/kg = 4.1868 kJ/kg = 4.1868 · 10³ J/kg.
 1 erg/g = 10⁻⁴ J/(kg·K).
Thermal conductivity. 1 J/(cm·s·K) = 1 W/(cm·K) = 100 J/(m·s·K) =
 = 100 W/(m·K).
 1 cal/(cm·s·K) = 418.68 W/(m·K).
 1 cal/(m·s·K) = 4.1868 W/(m·K).
 1 kcal/(m·h·K) = 1.163 W/(m·K).
Temperature conductivity. 1 m²/h = 2.7778 · 10⁻⁴ m²/s.
Heat transfer coefficient. 1 kcal/(m²·K) = 1.163 W/(m²·K).
Quantity of electricity. 1 coulomb = 1 C = 1 A·s = 1 J/V.
 1 ampere-hour = 1 A·h = 3.6 kC = 3600 C.
Current. 1 ampere = 1 A = 1 C/s = 1 V/ohm = 1 V·mho = 1 W/V.
Current density. 1 A/mm² = 1 MA/m² = 10⁶ A/m².
 1 A/cm² = 10 kA/m² = 10⁴ A/m².
Voltage. 1 volt = 1 V = 1 A·ohm = 1 A/mho = 1 J/C.
Electric field intensity, electric strength.
 1 V/cm = 100 V/m.
 1 kV/mm = 1 V/μm = 1 MV/m = 10⁶ V/m.
 1 volt per mil = 1 kilovolt per inch = 39.36 V/mm = 39.36 kV/m =
 = 3.936 · 10⁷ V/m.
Capacitance. 1 farad = 1 F = 1 C/V = 9 · 10¹¹ cm.
 1 centimetre of capacitance = 1 cm = 1.111 pF = 1.111 · 10⁻¹² F.
Resistance. 1 ohm = 1 Ω = 1 mho⁻¹ = 1 V/A = 1 V²/W.
Conductance. 1 siemens = 1 mho = 1 ohm⁻¹ = 1 A/V = 1 W/V².
Volume resistivity.
 1 ohm-centimetre = 1 ohm·cm = 0.01 ohm·m = 1.111 ps =
 = 1.111 · 10⁻¹² s.
 1 ohm·mm²/m = 10⁻⁴ ohm·cm = 1 μ ohm·m = 10⁻⁶ ohm·m.
 1 s = 9 · 10¹¹ ohm·cm = 9 · 10⁹ ohm·m.
Volume conductivity.
 1 ohm⁻¹·cm⁻¹ = 1 mho/cm = 100 ohm⁻¹·m⁻¹ = 100 mho/m = 9 · 10¹¹ s.
 1 s⁻¹ = 1.111 pmho/m = 0.1111 nmho/m = 1.111 · 10⁻¹⁰ mho/m.
Electric moment. 1 debye = 1 D = 10⁻¹⁸ cgs units of electric
 moment = 3.333 · 10⁻³⁰ C·m.
Magnetic flux. 1 maxwell = 1 Mx = 10 n Wb = 10⁻⁸ Wb.
Magnetic induction. 1 gauss = 1 Gs = 0.1 mT = 10⁻⁴ T.
Magnetomotive force. 1 gilbert = 1 Gb = 10/(4π) A ≈ 0.8 A.
Magnetic field intensity. 1 oersted = 1 Oe = 10³/(4π) A/m =
 = 79.577 A/m ≈ 80 A/m.
Luminance. SI unit is candela per square metre (the former name
 of this unit is "nit"—nt) = 1 cd/m².
Quantities in the field of ionizing radiations. SI units: for radiation
 dose—J/kg, for exposure dose of X-ray and gamma-radiation—C/kg,

for the power of exposure dose of X-ray and gamma-radiation—A/kg, for radiation intensity—W/m², for the activity of nuclide in a radioactive source—s⁻¹, for the density of flux of ionizing particles or photons—s⁻¹.m⁻².

1 rad (not to be taken as "rad" for a radian)=0.01 J/kg.

1 rad per second=1 rad/s=0.01 W/kg.

1 curie=3.7·10⁴ rutherfords=3.7·10¹⁰s⁻¹.

1 rutherford=10⁶ s⁻¹.

1 roentgen=1 R=2.57976·10⁻⁴ C/kg.

1 roentgen per second=1 R/s=2.57976·10⁻⁴ A/kg.

1 erg per second per square centimetre=1 erg/(s·cm²)=1mW/m²=
=10⁻³ W/m².

Amplification. Amplification, attenuation, etc., are estimated by such dimensionless units as bel (B), decibel (dB) and neper (Np).

The ratio between a bel, decibel and neper is:

$$1 \text{ dB} = 0.1 \text{ B} = 0.1151 \text{ Np}$$

$$1 \text{ Nep} = 8.686 \text{ dB} = 0.8686 \text{ B}$$

INDEX

- Acceptor, 343
- AC-circuit, 26-91
- Activation, 287
- Alloyed junction, 347
- Amplification factor, 309, 330
- Amplifier, 24, 372
 - amplification stage, 373
 - bandwidth, 379
 - bandpass, 584, 607
 - class A, 409, 412, 436
 - class AB, 415
 - class B, 415, 436
 - class C, 436
 - current gain, 374
 - efficiency, 409
 - equivalent circuit, 322, 387
 - frequency distortion in, 378
 - frequency response, 378
 - input admittance, 383, 388
 - input capacitance, 388
 - input resistance, 372
 - narrow-band, 377
 - output admittance, 384
 - output stage, 373
 - passband, 376
 - power gain, 324, 373
 - preamplifier stage, 373
 - primary-signal, 376
 - push-pull, 410, 412
 - radio-signal, 376
 - resistance-coupled, 378, 389, 396
 - transfer admittance, 384
 - transformer-coupled, 378, 416
 - tuned, 377, 584
 - sensitivity, 379
 - voltage gain, 374, 387
 - wide-band, 377
 - y-parameters, 385
- Amplitron, 507
- Anion, 342
- Anode, 280, 283
- Anode maximum dissipation, 298
- Antenna, 14
 - antiphased, 171, 195
 - asymmetrical dipole, 171
 - cage, 221
 - Chireix-Mesny, 210
 - counterpoise, 192
 - dielectric, 231
 - directional, 187
 - discone, 230
 - dominant wave in, 192
 - double-dipole, 171-72
 - effective height, 183
 - efficiency, 174, 215
 - folded-dipole, 171, 210
 - gain, 187
 - half-wave dipole, 174
 - harmonic, 195
 - higher harmonics in, 192
 - horizontal, 203, 207
 - horn, 228
 - image, 189
 - inphased, 171, 195
 - long-wave, 214-17
 - loop, 212
 - L-shaped, 171, 216
 - magnetic, 593
 - medium-wave, 214-17
 - metal-plate lens, 232
 - parallel resonance in, 179
 - Pistolkors', 171, 220
 - radiating properties, 175
 - reciprocity of, 187
 - resonance properties, 180
 - rhombic, 223
 - series resonance in, 178
 - short-wave, 217-25
 - slot, 227, 230

- tapped horizontal, 171
- tilted, 171
- travelling-wave, 223
- T-shaped, 171, 216
- vertical, 207
- V-shaped, 169, 220
- ultrashort-wave, 225-32
- wave impedance, 178
- Atmospherics, 259, 612
- Attenuation (along path), 273
- Audio amplifier, 403-15
 - feedback in, 417
- Automatic gain control, 650
- Axial-mode radiator, 197
- Backscattering effect, 264
- Backward-wave tube, 492-96, 502
- Balanced modulator, 552
- Barrier region, 345
- Bias, automatic, 446, 453, 461
 - cathode, 447
- Bias voltage, 430
- Blocking oscillator, 528-31
- Bonding electron, 340
- Breakdown voltage, 229
- Bunch, 484, 491, 495
- Buncher, 484
- Capacitance, input, 314
 - interelectrode, 314
 - output, 314
 - transfer, 314
- Capacitor, 23, 31
 - gang, 585
- Carrier frequency, 16, 377, 536
- Catcher, 484
- Cathode, 280, 283, 287
 - activation, 287
 - carbonized, 287
 - directly heated, 288, 305
 - impregnated, 289
 - indirectly heated, 288
 - oxide-coated, 288, 305
 - thoriated, 288
 - tungsten, 287, 305
- Cathode follower, 421, 642
- Cathode grid, 335
- Cathode-ray oscilloscope, 366
- Cathode-ray tube, 365
 - brightness, 368
 - control electrode, 366
 - electron gun, 366
 - electrostatic, 365-66
 - focusing coil, 371
 - linear scanning, 368
 - magnetic deflection, 366
 - saw-tooth pulse, 368
 - scanning voltage, 368
 - secondary emission in, 367
 - sensitivity, 368
 - sweep generator, 369
- Cathode-ray tuning indicator, 648
- Cavity resonator, 482-87
- Characteristic impedance, 112, 127
- Charge carrier, 341
- Coaxial cable, 148
- Collector, 283
- Collector stabilization, 392
- Comb, 492
- Common-base circuit, 410
- Common-emitter circuit, 410, 415
- Communication, 15
 - multichannel, 557
 - satellite, 275
 - tropospheric, 271
- Communication system, 17
- Common-emitter transistor, 359
 - current gain, 362
 - input characteristic of, 360
- Common-emitter transistor, 359
 - open-base emitter current in, 360
 - output characteristic of, 360
 - power gain, 362
 - voltage gain, 362
- Compensation, high-frequency, 427
- Complex quantity, 42
- Conduction band, 337-38
- Conduction electron, 340
- Conduction intrinsic, 340
- Counter-reflector, 226
- Coupled circuit, 92-105
 - insertion loss in, 94
 - lower resonance frequency, 100
 - upper resonance frequency, 100
- Coupling, 92
 - direct inductive, 103
 - external capacitive, 104
 - internal capacitive, 104
 - mutual inductance, 93
 - optimal, 105
 - resistive, 104

- Coupling coefficient, 99
- Coupling frequency, 97
- Covalent electron, 340
- Critical frequency, 249, 257
- Crystal diode, 337, 346
 - characteristic curve of, 348
 - forward resistance, 348
 - rectified current resistance, 348
- Crystal lattice, 339
- Crystal structure, 339
- Current, alternating, 26
 - convection, 12
 - nonsinusoidal, 55
- Cutoff frequency, 109, 113
- Cutoff voltage, 302
- Cutoff wavelength, 156

- Decibel, 110, 375
- Depletion region, 345
- Detection, 376, 561
- Detector, 17, 561, 580, 633
 - envelope, 642
 - full-wave, 639
 - ratio, 645
- Detuning, 70
- Device, active, 301
 - electron vacuum, 278
 - matching, 161
 - non-linear, 294
 - passive, 301
 - semiconductor, 278
- Dielectric, 338
- Dielectric losses, 146, 149
- Differentiating circuit, 575
- Diffraction, 245
- Diode, 285
 - AC resistance, 296
 - anode-cathode capacitance, 298
 - anode maximum dissipation, 298
 - breakdown voltage, 299
 - characteristic of, 292
 - DC resistance, 295
 - double, 336
 - internal resistance, 294
 - peak inverse anode voltage, 299
 - service life, 300
 - slope of characteristic, 279
- Diode-pentode, 336
- Diode-triode, 336
- Dipole, asymmetrical, 171
 - bent, 210
 - complex, 207
 - folded, 210
 - double, 171-72
 - half-wave, 174
 - magnetic, 227
 - Nadenenko, 221
 - quarter-wave, 190
- Directivity, 186
- Director, 202
- Discriminator, 643
- Distortion, 637-38
 - non-linear, 296
- Distributed element system, 125
- Diversity reception, 261
- Donor, 342
- Doping, 341, 352
- Double-electric layer, 284
- Drift space, 484
- Dynatron effect, 330
- Dynode, 335

- Echo, 260
 - round-trip, 260
- Effective resistance, 65
- Effective value, 45
- Electrical length, 135
- Electron current, 340
- Electron density, 341
- Electron-hole pair, 341
- Electronic efficiency, 437
- Electronics, molecular, 278
 - quantum, 278
- Electronic tuning, 487
- Electrostatic induction, 432
- Electron vacuum device, 278
- Electron valve, 279
 - envelope, 279, 285
- Electron-volt, 282
- Emission, 284
 - photoelectric, 285
 - secondary, 285
 - thermionic, 285
- Emitter, 287
- Emitter follower, 421, 642
- Emitter stabilization, 393, 410
- Energy gap, 338

- Energy level, 337
- Envelope, AM-signal, 536
- Equation of motion (of electron), 281
- Exponential law, 52
- Fading, 254
- Feedback, 325
 - in AF amplifiers, 417
 - negative, 418
- Feedback loop, 418, 449
- Feedback ratio, 419, 450
- Fidelity, 379, 584
- Filament, 286
- Filter, 24, 106
 - amplitude-frequency response of, 110-11, 114-15
 - attenuation factor, 110
 - band-elimination, 106, 117
 - bandpass, 106, 117
 - decoupling, 389
 - high-pass, 106, 114, 404
 - low-pass, 106, 109, 404
 - L-section, 107
 - T-section, 108
 - II-section, 107
- Forbidden band, 338
- Forced oscillation, 66
- Form factor, 590
- Foule-Lenz law, 174
- Frame grid, 335
- Frequency, 26
- Frequency band, 237
- Frequency conversion, 336
- Frequency convertor, 565-70, 618
- Frequency distortion, 378
- Frequency divider, 569
- Frequency multiplier, 439, 565
- Frequency pulling, 513-17
- Frequency response, 378
- Fundamental component, 55
- Fundamental frequency, 380
- Gain, 318
- Gang capacitor, 585
- Gang tuning, 587
- Generator, 24
- Grid, 302
- Grid bias, 315
- Grid current, 303
- Grid-leak, 395, 397
- Graphical method, amplifier, 382
- Ground wave, 18, 238, 254, 262
- Group velocity, 156
- Harmonic, higher, 55
- Heater, 288
- Hole, 340
- Hole current, 340
- Hole density, 341
- Hole mobility, 341
- Horn, biconical, 229
 - conical, 229
 - pyramidal, 228
 - sectoral, 228
- Hum, 288, 413
- Huygens principal, 245
- Hybrid circuit, 364
- Ideal transmission formula, 234
- Image charge, 188
- Impedance, 40
- Impurity, 341
- Incident wave, 129, 131
- Induced charge, 432
- Induced current, 432
- Inductance, mutual, 93
- Induction coil, 23, 31
- Injection synchronization, 517
- Insertion loss, 94
- Integrated circuit, 364, 577
- Interdigital structure, 503
- Interelectrode capacitance, 314
- Interference, 260
 - atmospheric, 19
 - industrial, 19
 - man-made, 612, 632
- Internal resistance, signal source, 315
- Intrinsic conduction, 340
- Ionization, 240-41
 - density, 240, 248
- Ionosphere, disturbances in, 243, 260
 - layers, 241
 - lower, 240
 - upper, 240, 243

- Ionospheric disturbance, 243, 260
Ionospheric storm, 244
- Junction diode, 347
 germanium, 347
 volt-ampere characteristic of, 347
- Junction potential difference, 344
- Keying, 560
- Klystron, 481, 484
 reflex, 485
- Leakage, 146
- Lengthening coil, 214
- Lighthouse triode, 474
- Limiter, 571, 644
 peak, 573
- Line of sight, 24
- Line-of-sight transmission, 238, 265
- Load line, 320, 334
- Lobe, 203, 208
 main, 228
 side, 228
- Local oscillator, 336, 619
- Lock-in, 517-23
- Loop, 130, 141, 162
- Loss tangent, 40
- Loudspeaker, 403
- Lower frequency, 465
- Lumped-element system, 125
- Magic eye, 648
- Magnetron, 503
 mode skipping in, 506
- Match, 142
- Matching, 112
- Matching device, 161
- Maximum anode dissipation, 407
- Maxwell's electromagnetic theory, 13
- Microminiaturization, 278, 363
- Mismatch, 114, 141
- Mixer, 558, 619
- Mode skipping, 506
- Modulation, 16, 376, 533
 amplitude, 533
 anode, 541-43
 anode-screen, 543-44
 frequency, 533, 545
 grid, 539
 keying, 560
 phase, 533, 545
 pulse, 555
 pulse amplitude, 557
 pulse duration, 557-58
 pulse position, 559
 screen-grid, 543-44
 suppressor-grid, 543-44
- Modulation factor (depth), 535, 539
- Modulation index, 546
- Modulator, 16
 balanced, 552
- M-type oscillator, 496-507
- Multiplex method, 266
- Multivibrator, 524-28
- Node, 130, 141, 162
- Noise background, 262
- Non-linear distortion, 380
 factor, 381
- Omnidirectional radiator, 203
- Operating point, amplifier, 382
- Oscillator, 431
 amplitude-balance equation, 454
 autotransformer-feedback, 460
 Colpitts, 459
 common-anode, 465
 common-cathode, 464
 common-grid, 464, 471
 cutoff, 441
 dynamic characteristic curve of, 440
 efficiency, 445
 electron-coupled, 463
 frequency pulling, 513-17
 Hartley, 459
 injection synchronization in, 517
 load characteristic of, 443-44
 lock-in, 517-23
 M-type, 496-507
 negative-resistance, 510
 O-type, 496

- overload operation, 441
- phase-balance condition, 450, 467
- power gain, 433
- push-pull, 477
- RC-type, 512
- relaxation, 524
- self-excitation, 450
- self-excited, 434, 449
- self-oscillation in, 450
- separately-excited, 434
- transitron, 511
- three-point configuration, 457
- underload operation, 442-44
- Oscillatory circuit, 23, 64
 - damping factor, 82
 - free oscillation in, 64
 - effective resistance, 65
 - parallel, 72, 77
 - series, 79
 - tapping factor, 88
 - transient process in, 81
- O-type oscillator, 496
- Outer-shell electron, 283
- Pair generation, 341
- Parent atom, 343
- Parent material, 341
- Parallel feed, 447, 460
- Passband, 71
- Passive network, 374
- Path attenuation, 273
- Pentode, 333
 - transconductance, 333
- Period, 26
- Phase, 26
- Phase deviation, 545
- Phase distortion, 380
- Phase shifter, 549
- Phase velocity, 153
- Platinotron, 507
- Point-contact diode, 349
- P-N junction, 343
- Polarization, 184
- Popov's receiver, 21
- Popov's transmitter, 21
- Potential barrier, 284, 292, 344
- Power gain, 324, 438
- Power, reactive, 45, 66
- Preamplifier, 373
- Preselector, 619
- Primary electron, 331
- Propagation, 24, 256
 - one-hop, 260
 - two-hop, 256, 260
- Propagation velocity, 14
- Q-factor, 41, 66, 71
- Quiescent point, amplifier, 382
- Radiation belt, 244
- Radiation losses, 161, 174
- Radiation pattern, 185
- Radiation resistance, 174
- Radio-relay system, 266
- Ratio detector, 645
- Reactance, capacitive, 36
 - inductive, 38
 - mutual, 93
- Reactive stub, 163
- Receiver, 17
 - adjustments in, 647-53
 - bandwidth, 590
 - controls in, 647-53
 - crystal, 579
 - direct-detection, 585
 - fidelity, 584
 - selectivity, 580
 - sensitivity, 582
 - superheterodyne, 585
- Recombination, 341
- Rectifier, full-wave, 301
 - half-wave, 301
- Reflected wave, 18, 19, 139
- Reflection coefficient, 139, 143
- Reflector, 200, 225
 - active, 201
 - focus of, 225
 - passive, 201
- Refraction, 247
- Refractive index, 247, 270
- Repeater, 266
- Resistor, 23, 31
- Resonance, common, 104
 - current, 74
 - first particular, 104
 - full, 105
 - parallel, 74, 178
 - second particular, 104

- series, 69, 178
- voltage, 69
- Resonance frequency, 69
- Resonator, 482
 - cavity, 482-87
 - vane-type, 507
- Retransmitter, 274
- Saturation, 290
- Saturation current, 291
- Saturation voltage, 291
- Scatter, 245
- Scattered reflection, 245, 270
- Schottky effect, 294
- Screen grid, 325
- Secondary electron, 331
- Secondary emission, 331
- Selectivity, 72
- Self-induction, 29
- Self-oscillation, 450
- Semiconductor, 24
 - barrier region, 345
 - charge carrier, 341
 - depletion region, 345
 - doping, 341
 - electron current, 340
 - equilibrium in, 344
 - hole mobility, 341
 - majority carrier, 342
 - minority carrier, 342
 - N-type, 342
 - pair generation, 341
 - P-N junction, 343
 - potential barrier, 344
 - P-type, 343
 - recombination, 341
- Semiconductor device, 278
- Series feed, 447
- Service life, 300
- Shortening capacitor, 215
- Sideband, 536
- Signal conversion, 532
- Skin effect, 31
- Skip distance, 257
- Sky wave, 18, 252, 254, 262
- Slow-wave structure, 490
- Solid circuit, 364
- Sounding, oblique, 265
 - vertical, 264
- Space charge, 289, 308, 322
- Spectrum, 57
 - amplitude-frequency, 63
 - phase-frequency, 63
- Standing wave, 130, 133, 136, 141
- Standing wave ratio, 142
- Static(s), 259
- Steady-state condition, 54
- Substitution electron, 340
- Suppressor grid, 333
- Sweep generator, 369
- Temperature drift, 392
- Tetrode, 325, 543
 - amplification factor, 330
 - beam, 331-32
 - characteristic curve of, 330
 - dynatron effect in, 330
 - secondary emission in, 331
 - transfer capacitance, 331
- Thin-film circuit, 364
- Three-halves power law, 292
- Time constant, 51
- Transconductance, 310, 333, 384
 - dynamic, 319
- Transfer capacitance, 326
- Transformation ratio, 405-6, 417
- Transformer, impedance, 161, 163
 - matching, 225
 - quarter-wave, 166
 - tuned, 161
- Transient process, 50
- Transit time, electron, 281
- Transmission distance, 539
- Transmission line, 120-68
 - aerial, 147
 - asymmetrical, 148
 - electrical length, 135
 - finite length, 128
 - four-wire, 150
 - ideal balanced two-wire, 120
 - incident wave in, 129, 131
 - loss in, 146
 - lossless, 120
 - matching of, 138
 - reflected wave in, 129, 139
 - reflection coefficient, 139, 143
 - wave impedance, 127
- Transmitter, 17
- Transistor, 24, 337, 351
 - base, 352

- base current, 353
- collector, 352
- collector current, 353, 357
- collector cutoff current, 354, 356
- common-base, 354, 356
 - input characteristic of, 355
 - output characteristic of, 355
- emitter, 352
- emitter current, 353
- forward voltage, 352
- N-P-N type, 352
- P-N-P type, 352
- reverse voltage, 352
- Transistor amplifier, 357
 - collector stabilization in, 392
 - common-base current gain, 358
 - current gain, 358
 - emitter stabilization in, 393, 410
 - input resistance, 357
 - line of load, 358
 - output resistance, 357
 - power gain, 358
 - temperature drift in, 392
 - temperature stability of, 392
 - voltage gain, 358
- Travelling wave, 123, 141
 - equation, 124
 - ratio, 141
 - tude, 488
- Trigger, 573
- Triode, 302
 - AC anode resistance, 311
 - amplification factor, 309
 - anode characteristic of, 308
 - anode-grid characteristic of, 304
 - cutoff voltage, 302
 - grid characteristic of, 306
 - dynamic characteristic of, 319
 - family of characteristics, 307
 - grid current, 303
 - maximum anode dissipation, 308
 - metal-ceramic, 474, 480
 - space charge in, 308
 - static characteristic of, 308, 319
 - transconductance, 310
 - twin, 303
- Triode-heptode, 336
- Triode-pentode, 336
- Troposphere, 239
- Tuned circuit, 541
- Tuning, ganged, 587
- Tunnel diode, 350
- Tunnelling effect, 350
- Turbulence, 239
- Upper frequency, 465
- Valence band, 337-38
- Valence electron, 337
- Valve, 23
 - dual-control, 336
- Vane-type resonator, 507
- Varicap, 349
- Vector diagram, 28
- Velocity modulation, 484, 489
- Video amplifier, 421-27
- Voice coil, 403
- Voltage divider, 393
- Voltage gain, 322
- Wave, AM modulated, 16
 - dominant magnetic, 159
 - electromagnetic, 11-13
 - higher order, 160
- Wave
 - long, 19
 - medium, 19
 - short, 19
 - transverse electrical, 159
- Wave front, 152, 235
 - spherical, 234
- Waveguide, 151
 - circular, 151, 160
 - rectangular, 151
 - tropospheric, 270
- Wave impedance, 112, 127
- Wavelength, 15, 123
- Work function, 285
- Y-parameters, 385
- Zone of silence, 261

TO THE READER

Mir Publishers welcome your comments on the content, translation, and design of the book.

We would also be pleased to receive any suggestions you care to make about our future publications.

Our address is:

USSR, 129820, Moscow, I-110, GSP, Pervy Rizhsky Pereulok, 2, Mir Publishers.